Inferring organismal and character evolution from functional genome features

Von der Fakultät für Lebenswissenschaften

der Technischen Universität Carolo-Wilhelmina zu Braunschweig

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

genehmigte

D i s s e r t a t i o n

von Palani Kannan Kandavel
aus Virudhunagar, Indien

1. Referent: Prof. Hans-Peter Klenk
2. Referent: Prof. Dietmar Schomburg
eingereicht am: 03.11.2014
mündliche Prüfung (Disputation) am: 05.03.2015
Druckjahr 2015

Vorveröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Fakultät für Lebenswissenschaften, vertreten durch **den Mentor** der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Publikationen

1. Pagani, I., Lapidus, A., Nolan, M., Lucas, S., Hammon, N., Deshpande, S., Cheng, J.-F., Chertkov, O., Davenport, K., Tapia, R., Han, C., Goodwin, L., Pitluck, S., Liolios, K., Mavromatis, K., Ivanova, N., Mikhailova, N., Pati, A., Chen, A., Palaniappan, K., Land, M., Hauser, L., Chang, Y.-J., Jeffries, C.D., Detter, J.C., Brambilla, E., **K. Palani Kannan**, Ngatchou Djao, O.D., Rohde, M., Pukall, R., Spring, S., Göker, M., Sikorski, J., Woyke, T., Bristow, J., Eisen, J.A., Markowitz, V., Hugenholtz, P., Kyrpides, N.C. and Klenk, H.-P. "Complete genome sequence of *Desulfobulbus propionicus* type strain (1pr3^{T})." *Standards in Genomic Sciences* 4.1 (2011): 100-110.
2. Pati, A., Abt, B., Teshima, H., Nolan, M., Lapidus, A., Lucas, S., Hammon, N., Deshpande, S., Cheng, J.-F., Tapia, R., Han, C., Goodwin, L., Pitluck, S., Liolios, K., Pagani, I., Mavromatis, K., Ovchinikova, G., Chen, A., Palaniappan, K., Land, M., Hauser, L., Jeffries, C.D., Detter, J.C., Brambilla, E.-M., **K. Palani Kannan**, Rohde, M., Spring, S., Göker, M., Woyke, T., Bristow, J., Eisen, J.A., Markowitz, V., Hugenholtz, P., Kyrpides, N.C., Klenk, H.-P. and Ivanova1, N. "Complete genome sequence of *Cellulophaga lytica* type strain (LIM-21^{T})." *Standards in Genomic Sciences* 4.2 (2011): 221-232.
3. Ivanova, N., Sikorski, J., Chertkov, O., Nolan, M., Lucas, S., Hammon, N., Deshpande, S., Cheng, J.-F., Tapia, R., Han, C., Goodwin, L., Pitluck, S., Huntemann, M., Liolios, K., Pagani, I., Mavromatis, K., Ovchinikova, G., Pati, A., Chen, A., Palaniappan, K., Land, M., Hauser, L., Brambilla, E.-M., **K. Palani Kannan**, Rohde, M., Tindall, B.J., Göker, M., Detter, J.C., Woyke, T., Bristow, J., Eisen, J.A., Markowitz, V., Hugenholtz, P., Kyrpides, N.C., Klenk, H.-P. and Lapidus, A. "Complete genome sequence of the extremely halophilic *Halanaerobium praevalens* type strain (GSL^{T})." *Standards in Genomic Sciences* 4.3 (2011): 312-321.

Posterbeiträge

1. **K. Palani Kannan**, Göker, M. and Klenk, H.-P. “Inferring functional genome classifications from their annotated proteins.” *German Conference on Bioinformatics 2011* (2011).

ACKNOWLEDGEMENTS

I would like to thank my supervisors, PD Dr. Markus Göker and Prof. Dr. Hans-Peter Klenk, for encouraging my research and for allowing me to grow as a researcher. Your advices and help in integrating myself in the scientific society is priceless.

I thank all my thesis committee members Prof. Dr. Dietmar Schomburg, Prof. Dr. Ralf Rabus and PD Dr. Joern Petersen for their constructive criticism and valuable suggestions.

I would like to thank Prof. Dr. Joerg Overmann, Managing Director of DSMZ, Ms. Bettina Fischer, Head of Administration and the management members of DSMZ to provide me an opportunity and facilities to work in DSMZ.

I would like to thank Prof. Dr. Meinhard Simon and Collaborative Research Centre Transregio 51 for supporting me financially for my PhD work.

I am extremely grateful to my colleagues Dr. Jan Meier-Kolthoff, Carmen Scheuner and Dr. Maria del Carmen Montero-Calasanz for their valuable suggestions and collaborative work.

I thank Prof. Dr. Dietmar Schomburg and his team for providing me the data which can be used to evaluate and apply my methodology.

I remind of Prof. Dr. Dietmar Schomburg and Prof. Dr. Miguel Vences for their help to recognize my masters degree in TU-Braunschweig.

I thank Dr. Birte Junge, Dr. Tanja Piekarski and Dr. Ferdinand Esser. They gave me many opportunities in extra-curricular exposure through Roseobacter Grad School.

I wish to thank all DSMZlers for including me as one among them and their friendliness. With their support, I proceeded my work successfully.

Finally, I am grateful to all my international friends, International Students Network, (TU-Braunschweig) and Internationale Gauss-Freunde (TU-Braunschweig) for their support during my stay in Braunschweig.

I am very much thankful to my teachers, lecturers, professors and family members for their encouragement on my work.

Table of Contents

SUMMARY

The development of phylogenomics pipelines is important as phylogeny-driven genome sequencing projects generate plenty of genomic data. This thesis work focused on the development of three pipelines which yield: 1. proper taxonomic descriptions of new genomes; 2. phylogeny using genome encoded functional features; and 3. evolutionary correlations between functional characters (e.g., genes), environmental characteristics, genome features of newly sequenced genomes and their functional linkages.

A contemporary way of inferring phylogenies using genome encoded COGs, pathways and GOs are optimized. 34 variations of distance-based phylogenetic tree reconstruction strategies for eight datasets were formulated with regard to similarity calculation methods, data sources, data observation strategies (presence/absence and abundance), weights, and filters applied to genome encoded functionalities and COGs. The suitability of these distance formulae for phylogeny reconstruction were estimated by computing treelikeness from the distance matrices. The phylogenetic reconstruction methods which have elevated treelikeness among eight datasets were identified and reported with the help of multiple-regression analysis. Robinson-Foulds distances between phylogenetic trees created using 34 variations mentioned above and sequence-based trees were calculated. The COG-based trees that are topologically similar with ortholog/gene-content based trees were reported with the help of the Principal Coordinates Analysis. The distance matrices created using COGs obtained from IMG have high treelikeness and are congruent with previously established sequence-based trees. From the comparison of strategies, the optimal strategy was found with the formulation that contains: 1. functionalities mapped from IMG COGs and IMG COGs obtained from IMG as data source; 2. presence/absence method as best strategies when compare to abundance based strategies; 3. complete presence (weighted with 1.0) as a best threshold value for calculating presence of functionalities; and 4. sorensen's method as similarity calculation method. From this results, it has been learned that COG-based phylogenetic reconstruction strategy can be used as an alternative for ortholog-content based strategies. The future clade-specific comparison between pathway/GO-based and sequence-based trees can help to infer the complete metabolic evolution and biochemical characteristics evolution across sets of organisms.

A pipeline which calculates the evolutionary correlation between functional characters, genomic features and their functional linkages is developed as well as optimized. BayesTraits software was used

in the pipeline to estimate the correlated evolution between character pairs. Characters were clustered using the MCL algorithm with regard to significant evolutionary correlations between characters. The effect of small genomes with genes lost in parallel was identified and eliminated. The evolutionary correlation of genes grouped in COG groups/COG categories and mapped to the same pathway were statistically interpreted for a *Spirochaetae* dataset which contains 29 genomes with the help of chi-squared test. The genus *Spirochaetae* contains diverse motile species, thus evolutionary correlation between genes mapped to motility pathways of *Spirochaetae* genomes were cross-verified with previous experimental evidence. The study provides the following findings: 1. the number of evolutionary correlated genes in motility pathways of the *Spirochaetae* dataset were high; and 2. the evolutionary gain/loss of motility genes in *Spirochaetae* spp. were reported. The evolutionary correlation between enzymes mapped to the same pathways of a *Rhodobacteraceae* dataset which contains 87 genomes were analysed and its applications were explained with an example. The study provides the following key finding as 25% of the pathways were completely mapped with enzymes evolved in correlated manner. The pathways with evolutionary correlations were identified and an example of chloroaromatic hydrocarbon degradation pathways were listed and explained with their functional relations for the reason that several *Rhodobacteraceae* spp. degrade hydrocarbons in the environment. The genomic features which correlated in evolution were identified and reported. The high capability of pathogens to gain cp4-44 prophage element in *E. coli* + *Shigella* dataset was identified. The similar character history trace and correlation of (S)-2-haloacid dehalogenase with marine living environmental characteristics in *Rhodobacteraceae* dataset was reported. The disimilar character histroy trace and correlation of ectoine biosynthesis pathway in 34 marine organisms were described. The correlation between 2,3-dihydroxy benzoate biosynthesis pathway and non-marine Paracoccuss spp. was identified. The importance of newly developed strategy has been explained with the uncorrelated characters with their character history traces (Glycogen biosynthesis pathway I with marine living environmental characteristics of *Rhodobacteraceae)*.

1. INTRODUCTION

As many scientific methods in biological disciplines are developing rapidly, heterogeneous biological data are increasingly produced and available as scattered data sources. For instance, genomic data is one of the rapidly growing biological data [Pagani et al., 2012]. The genomic data generated by various genome sequencing centers have been rapidly increasing since the year 2000 [Benson et al, 2000]. The major genomic data are available in public repositories collaborated under INSDC (International Nucleotide Sequence Database Collaboration) such as Genbank [Luscombe et al., 2001]. Bioinformatics methods must be designed in such a way that they fulfill the needs of biologists when analyzing large genomic datasets through the use of sophisticated computational methods. [Kumar and Dudley, 2007]. Bioinformatics pipeline development plays a vital role in developing standardized genomic data analysis protocols. Pipeline development and system integration are two nascent bioinformatics approaches for the analysis and correlation of biological data. This thesis concentrates on the development of bioinformatics pipelines for the analysis and correlation of genomic data as means of simplifying the understanding of biological entities, particularly in the phylogenomics research. Three bioinformatics bottlenecks have been identified and pipelines have been implemented within the framework of this thesis. The three bottlenecks are: 1. a lack of comprehensive definitions regarding the taxonomy of microbial organisms; 2. a lack of standard genome classifications strategies based on their encoded functional functionalities; and 3. a lack of evolutionary studies on functionally linked genes/enzymes of newly sequenced genomes.

1.1. GENOME SEQUENCING

As a successful genome sequencing method, the sequencing procedure utilizing radio-labeled DNA polymerase was invented in 1975 under the name “Plus and Minus Technique” [Sanger, 1980]. In 1977, DNA sequencing based on the selective incorporation of chain-terminating dideoxynucleotides was invented by Sanger as the so-called “Sanger Sequencing Method” [Sanger et al., 1977]. This method was adopted by sequencing centers and successful for several decades. It is being used in current genome sequencing platform. However, next generation sequencing techniques such as pyrosequencing, illumina, heliscope, and AB SOLiD have emerged. Pyrosequencing and illumina techniques replaced the Sanger method in recent years [Shendure and Ji, 2008]. As advanced technologies have already emerged for genome sequencing, the sequencing technology will soon cease to be a limiting factor in this field [Scholz et al., 2012]. In this concern, genomic data have been rapidly

increasing. In total, 11,472 genome projects were started between 1997 and 2012. From this, 2907 genomes sequences have been completely sequenced as permanent drafts and 340 sequenced genome sequences are meta-genomes [Pagani et al., 2012].

1.2. GENOMICS

Genomics is a discipline that applies DNA sequencing methods and bioinformatics approaches to analyze the function, structure and evolution of genomes. Following the discovery of the structure of DNA by Watson and Crick, nucleotide sequencing became a major interest for molecular biologists [Ankey, 2003]. Advanced genome sequencing techniques led to the dynamic study of transcription, translation, protein-protein interactions, etc. From the completed genome sequence, details about an organism on the molecular level can be determined. For example, metagenomic studies (the study of genetic materials obtained directly from environmental samples) [Hugenholtz et al., 1998]. In this context, new bioinformatics approaches have been developed in order to understand and analyze large amounts of genomic data [Koonin, 2001].

The microbial (bacteria and archaea) domain constitutes large number of organisms [Cavalier-Smith, 1998]. New organisms are being identified from environmental samples. The genome sequencing strategies coupled with proper genome classifications are required as currently available genomes is limited by a highly biased phylogenetic distribution. This necessity galvanised researchers to launch genome projects for bacteria and archaea, such as the Genome Encyclopedia of Bacteria and Archaea (GEBA) [Wu et al., 2009]. In connection with these genome projects, the development of analogous bioinformatics approaches has become an essential requirement [Horner et al., 2010]. With several bioinformatics methods (e.g. genome assembly and gene annotation) already having been established in genome data analysis, this thesis focus on the development of bioinformatics methods for phylogenomics studies using microbial genomes.

1.3. PHYLOGENOMICS

Phylogenetics is a study comprising scientific approaches employed in order to gain insights into evolution. Until the 1970s, phylogenetic studies had been undertaken on the basis of morphological characters, ultra structural characters, biochemical and chemotaxonomical features. However, this approach is hampered by the limited number of reliable homologous characters available [Delsuc et al., 2005]. With the end of 1980s, access to DNA sequences increased the availability of large homologous characters for the inference of phylogeny, thereby greatly improving the resolving power of

phylogenetic inference. Oligonucleotide cataloging method had been used to define evolutionary relationships in bacteria [Delsuc et al., 2005]. In conjunction with the availability of homologous characters, phylogenetic reconstruction methods using molecular data were developed [Felsenstein, 1988]. Later, it had been replaced by 16S rRNA based evolutionary study, with the 16S rRNA being conserved [Woese, 1987]. As such, evolutionary studies based on 16S rDNA have become widely accepted in 1990s after the discovery of 16S rDNA amplification method [Weisburg et al., 1991].

In 1995, first complete genome of a bacteria had been sequenced [Fleischmann et al., 1995]. More genomic sequences were generated by large-scale genomic projects (e.g. GEBA was initiated in 2009). With the available genome sequences, this new field of research, termed phylogenomics, which uses phylogenetic principles to make sense of genomic data, received a sudden influx of data with which to process and analyze [Eisen and Fraser, 2003]. In a broad sense, phylogenomics involves understanding evolution through the use of genomic data. The term phylogenomics literally combines phylogenetics and genomics. The term was first introduced in 1998 to predict gene functions on the basis of their evolutionary information [Eisen, 1998]. The main foci in phylogenomics are to infer phylogenetic relationships using genomic data, to gain insights into the mechanisms of molecular evolution [Philippe and Blanchette, 2007], and to conduct multi-species comparisons in order to infer the evolution of molecular functions. Phylogenetic tree reconstruction methods are developed using various genomic data such as DNA sequences, complete gene-content, core gene-content and protein sequences. It extended research interests to include horizontal gene transfer, the loss/gain of molecular functions over the course of molecular evolution and taxonomical reclassification studies [Delsuc et al., 2005].

1.4. TAXONOMY

Carolus Linnaeus developed the Linnean classification system to classify biological organisms and name them according to binomial nomenclatures. Taxonomy is the science of grouping organisms based on their shared characteristics and naming those groups accordingly. Organisms are grouped according to their taxonomic ranking. Each organism has its own taxonomic designations with respect to its taxonomic ranks. Groups in a given rank can be further combined into higher, superordinate groups, thereby enabling, a taxonomic hierarchy to be created [Lapage et al., 1992].

The taxon name is an indication of taxon's characters or taxonomic classification which refers history of it. A set of rules has been established for defining a taxon name. Some of rules include the following critical points: 1. the name must be unique; 2. the name should be framed using up to only 26 letters

from the Latin; 3. the name must be assigned based on taxonomic hierarchy rules; 4. there should be enough scientific evidence to assign a specific taxonomic rank; and 5. the name should be published in permanent scientific literature. In most cases, the etymology of taxon names are derived from the geographic range of the taxon, ecological notes, chemistry, behavior, etc. In essence, it depends on the available data and resources [Lapage et al., 1992].

In the case of bacteria, taxon names are regulated on the basis of the International Code of Nomenclature of Bacteria (or Bacteriological Code) [Lapage et al., 1992], which has its own set of regulations. The taxon name is approved after it is published via any of the methods: 1. the name is cited in the approved list of bacterial names; 2. the name is published in an article in the International Journal of Systematic Bacteriology (IJSB) and conforms to the requirements set out in the bacteriological code; and 3. the name is published as an entry by announcement in a validation list published in the IJSB [Lapage et al., 1992]. In 2000, IJSB had been changed as International Journal of Systematic and Evolutionary Microbiology (IJSEM).

1.4.1. TAXONOMIC DESIGNATIONS

Even though the code of nomenclature is available, taxonomic designations and definitions are not always the same in long run for a specific organism over an extended period of time. Names can be revised for several reasons. For instance, if an alteration of the diagnostic characters or the circumscription of a taxon modifies the nature of the taxon itself, the taxon name of the organism will be emended [Euzeby, 2012]. The taxonomic designations and definitions are continually modified with respect to new scientific evidence gleaned for that particular organism. In conjunction with this, these changes should be properly identified and cited [Euzeby, 2012]. As genomic data are rapidly growing, an important requirement is to situate these extracted genomic data in a meaningful context that relates to proper taxonomic descriptions, as this would be highly useful to researchers in this field, especially if this information is readily available and accessible through a number of different channels [Garrity et al., 2009].

In chapter 1, a pipeline has been developed to show the details of various taxonomic designations and definitions of bacteria along with proper citations and etymologies. This is due to the fact that the proper taxonomic designation and definition of an organism which genome had been sequenced is a key prerequisite.

1.5. PHYLOGENETIC TREE RECONSTRUCTION

1.5.1. PHYLOGENETIC TREES

An evolutionary tree was first depicted by Charles Darwin in 1859 [Darwin, 1859]. Since then, evolutionary relationships have been illustrated in the form of trees or branching diagrams. A phylogeny is the evolutionary history of a group of organisms, and is normally depicted in terms of relative decency of common ancestry using various biological data and parameters. Each phylogenetic tree has its own nodes and branches, with the evolutionary relationships being represented as a branching diagram (or tree) in which branches are joined by nodes and lead to terminals at tips of the tree [Jill Harrison and Langdale, 2006]. An individual node represents the common ancestor shared by the branches extending from it. The trees can be divided into two types based on the representational direction of evolution. The types of trees are: 1. rooted tree (a directed tree with unique nodes corresponding to the recent common ancestor of all organisms depicted on the tree); and 2. unrooted tree (unique nodes which correspond without making an assumption of regarding the ancestry of the organisms).

The phylogenetic reconstruction rely on mathematical methods. The mathematical methods use shared homologous characters between different organisms to infer phylogenetic trees. The phylogenetic inference is strongly dependent on the mathematical method and data utilized. Thus, phylogenetic reconstruction methods which utilizes reliable biological data are being developed and successful. Theoretically, reliable data are understood to be those which have undergone only a few changes over time [Delsuc et al., 2005]. Genomic data comprises feature of reliable homologous characters (e.g. orthologous genes, house-keeping genes). It can be used in the reconstruction of phylogenetic trees. Thereby indicating that phylogenetic reconstruction using genomic data are reliable and available in public repositories [Delsuc et al., 2005, Sanderson et al., 2003].

1.5.2. PHYLOGENETIC TREE RECONSTRUCTION METHODS

Three major types of phylogenetic tree reconstruction methods, namely 1. Distance-based method; 2. Maximum parsimony; and 3. Maximum likelihood methods [Holder and Lewis, 2003].

A distance based method is a tree construction method using a matrix of pairwise distances between features of organisms. Neighbor joining is an example of a distance based tree construction method which employs the distance matrix. Maximum parsimony is a character-based tree construction method which uses discrete phylogenetic characters to infer one or more optimal phylogenetic trees for a group

of organisms. This method evaluates phylogenetic trees according to a precise optimality criterion. The tree with the most favorable score will be taken as the best estimate. Since several decades, maximum parsimony is a widely used character-based tree reconstruction method implemented in conjunction with morphological data. In maximum-likelihood method, the probability of possible tree outcomes using character data are computed. The highest likelihood tree is the preferred one. The probability of data given tree can be calculated from the set of probabilities of character changes. It is statistically well established method with good consistency of tree reconstruction and computationally intensive.

The phylogenomics based reconstruction methods use either sequence data, gene-content, ortholog-content or gene order data from whole genomes. Some of the strategies which use genomic data for phylogenetic reconstruction are explained below.

The so-called "genome BLAST distance phylogenies (GBDP)" uses whole genome data. GBDP has been developed based on pairwise similarities and genome distances obtained from the high-scoring segment pairs in BLAST analysis using sequences of plastids and mitochondrial genomes. This method is considered to be a statistically significant phylogenetic reconstruction method [Auch et al., 2006].

Phylogenomic studies using gene content started to be carried out from 2000 onwards [Baldauf et al., 2000]. As part of multi-gene-content based phylogenomics studies, a specific phylogenetic reconstruction study was initially conducted using 100 genes [Bapteste et al., 2002], which subsequently led to the possibility of using more than 100 genes in phylogenetic reconstruction methods [Driskell et al., 2004]. Following the principle of using all relevant genomic data for phylogenetic reconstruction, the most popular strategy is to reconstruct phylogeny using the "supermatrix" that is formed through the concatenation of genes [Delsuc et al., 2005]. The robustness of super matrix strategy has led to the powerful phylogenetic reconstruction method implemented today.

Phylogenetic relationships involving the presence/absence of orthologous genes were first studied by [Wolf et al., 2001]. As orthologous genes are vertically evolved genes, this method is less prone to homoplasy (genes evolved in a convergent manner). As such, orthologous genes are potential phylogenetic markers. The difficulties in this method is found to be the standard ortholog assessment procedure and the lack of evaluation [Delsuc et al., 2005].

The phylogenetic reconstruction studies have been carried out based on rare genomic changes (RGC) due to the fact that RGCs have a very low probability of being the result of convergence [Delsuc et al.,

2005]. Gene order, intron positions, insertions and deletions, retroposon integrations, gene fusion and fission events are described as RGCs [Rokas and Holland, 2000]. Only a few characters of this kind have been used to address phylogenetic questions [Delsuc et al., 2005].

1.5.3. TREE RECONSTRUCTION USING ENCODED FUNCTIONAL FUNCTIONALITIES OF GENOMES

Homoplasies and horizontal gene transfer between organisms are the pitfall in inferring phylogenetic trees. The correctness of inferences can be verified by obtaining confirmation from independent sources such as orthologous groups or functional groups. If phylogenetic inferences based on these independent sources converge with the relative results, increased confidence can be placed in the corresponding phylogeny [Philippe and Blanchette, 2007]. In chapter 2, phylogenetic trees are constructed using cluster of orthologous groups (COGs), gene ontologies (GOs) and pathways. A comparison of these trees and those which are already well established facilitates correction or improvement in the phylogenetic inference of groups of organisms.

Aside from phylogenetic tree reconstruction using COGs or GOs or pathways, the crucial part in the process is mapping them for newly sequenced genomes. A standard pipeline is required to map genes into corresponding COGs, GOs, and pathways. In chapter 2, a pipeline has been designed to map genes into corresponding genome encoded functional functionalities. Distance-based phylogenetic trees are created based on the presence/absence or abundance of COGs/GOs/pathways in a group of genomes.

1.6. EVOLUTION OF FUNCTIONALLY LINKED GENES

1.6.1. GENES

In classical genetics, a gene is a unit which transfers inheritance from the ancestral generation to the next generation. Nowadays, a gene is predominantly defined as a sequence of DNA which converts into strands of messenger RNA and could be used as the basis for building associated proteins/enzymes piece by piece [Pearson, 2006]. The sequence-based evolutionary study of genes in a genome-scale is challenging to conduct, it requires more powerful computational resources. In eukaryote, the abundance of conserved non-coding (intron) regions increases the complexity. If genes are considered as functional character instead of piece(s) of nucleotide sequence, the memory consumption in computational process will be reduced. Recently, the definition of gene has been generalized as “a union of sequences encoding a coherent set of potentially overlapping functional products” [Gerstein et al., 2007]. This definition has spurred researchers to explore the evolution of genes beyond the

sequence level. In this manner, a gene can be presented as a functional product rather than being considered it as a component in a whole-genome DNA sequence. The functional product may be a gene or protein or any specific functional characteristics derived from the genome [Gerstein et al., 2007].

1.6.2. FUNCTIONAL CLASSIFICATIONS

Functional classification is the process of classifying genes whilst taking account of the knowledge regarding various functional characteristics. A gene can be functionally classified into groups based on its involvement in metabolic activities, orthologous characteristics and functional characteristics, such as genes involved in a pathway or genes classified according to their molecular functions or cellular role.

Pathways are a series of chemical reactions catalyzed by enzymes. A collection of metabolic pathways is called a metabolic network. KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways [Kanehisa and Goto, 2000], MetaCyc [Caspi et al., 2008] and BRENDA [Schomburg et al., 2013] are some examples of resources for pathway information. Of the pathway resources available, KEGG pathway classification is the pathway resource used in thesis, as it is frequently updated to reflect newly sequenced genomes. KEGG pathway classification is a manually curated database which serves as a representation of knowledge about metabolic pathways and molecular networks. Each pathway can be visualized as a network of chemical reactions with catalyzing enzymes. In KEGG online resource, names of genes/proteins/enzymes can be mapped to corresponding pathways. This feature facilitates the grouping of genes according to their involvement in a pathway.

COG is a system of delineation for orthologous genes from sequenced genomes of prokaryotes and small eukaryotes [Tatusov et al., 2003]. The COG system is a database consisting of orthologous genes clustered according to sequence similarities. The elucidation of consistent patterns in sequence similarities allows for the delineation of 720 clusters in the first version [Tatusov et al., 1997]. In the recent version, delineations increased to 5665 clusters. Each cluster has a unique identification number. In effect, the COG database helps to bring comparative genomics and protein classification studies together. Of the numerous possible approaches to gene/protein classification, “the COGs appear to be unique as a prototype of a natural system, which has as its basic unit a group of descendants of a single ancestral gene” [Tatusov et al., 1997]. Each classification is associated with a specific conserved function, so that the inclusion of a protein in a COG helps to highlight the protein function [Tatusov et al., 1997]. The COG clusters are further divided into two levels featuring 22 COG categories and four

COG groups.

1.6.3. FUNCTIONAL GROUPS AND EVOLUTION

Transition and transversion in genetic materials are primary reasons for evolutionary changes. Models have been proposed for the study of the evolution of nucleotides/amino acid sequences. The Kimura two-parameter model [Kimura, 1980]and Dayhoff model are two examples [Dayhoff et al., 1978]. These models assist in the study of the rate of variation in sequences; in this context, the homologs and heterologs can be identified with regard to the variations in sequences. Homologs are classified into orthologs and paralogs. Orthologs are group of homologous gene/proteins which have diverged after a speciation event in the evolutionary process. Paralogs are groups of homologous gene/proteins which have diverged after a gene duplication event in the evolutionary process. In the essence, orthologs are genes evolved vertically, while paralogs are genes evolved in a non-vertical manner [Jensen, 2001]. With this in mind, it is possible to group genes and therefore their relationships in evolution can be studied.

The genes in the same molecular network or which share the same functional characteristics are called functionally linked genes. It is likely that the functionally linked genes will evolve together [Pellegrini et al., 1999]. The evolution of functionally linked genes can be studied with regard to gain and loss of genes across a range of genomes. Theories on the evolution of pathways are proposed as de novo invention, retro-evolution, enzyme recruitment, pathway duplication and specialization of multi functional enzymes [Schmidt et al., 2003]. Except the pathway duplication and de novo invention theories, all other theories agree that variations in enzyme structure/function affect pathway and cause variations in pathway [Schmidt et al., 2003]. As such, an evolutionary variation in functionally linked genes/enzymes in a pathway helps to understand the evolution of a whole pathway in a group of organisms [Pazos and Valencia, 2008]. The evolutionary correlation between functionally linked genes/proteins or functionally linked genomic features (e.g. length of a nucleotide sequence, number of genes encoded) have not been well explored in genome level. Furthermore, the evolutionary correlation between functionally linked genes can be used to study the adaptation of genes in a pathway which is responsible for the survival of an organism and changes in metabolic characteristics of an organism. For the reasons and requirements stated above, the evolutionary correlation of functionally linked genes, enzymes, pathways and their role in the molecular network are studied in the third chapter.

1.7. AIM OF THESIS

Genomic data are an excellent corpus for studying evolution. Phylogenomics plays a major role in connecting genomics and evolution with the aim to resolve issues in taxonomical classification. “There is an ever-growing list of examples in which cross-talk between these two disciplines (evolution & genomics together as phylogenomics) has enabled scientists to design better experiments and generate new insights” [Eisen and Fraser, 2003]. From the above motivations, the aim of thesis is therefore to resolve the bottlenecks (as mentioned in the Introduction) present in the field of phylogenomics by developing bioinformatics pipelines. In this thesis, newly sequenced bacterial genomes are used to validate the bioinformatics pipelines which have been developed. The three pipelines are described over the course of three chapters.

Chapter 1: The number of sequenced microbial genomes already reached 2907 and new sequencing projects are increasing rapidly since 10 years. However, the taxonomic history of sequenced organisms, their clear taxonomic nomenclatures and the definitions of organisms are not linked in the sequencing pipeline. The aim of the first chapter is to develop an application which comprehensively generates literally readable tabular and textual data comprising taxonomic and taxonomy-related details on microbes obtained from previously established local database which contains LPSN (List of Prokaryotic names with Standing in Nomenclature) data [Euzeby 2012]. A previously established database consists of taxonomic nomenclatures with citations, etymologies of taxon names, inventory citations about organisms, hyperlinks to the remote databases of rRNA data, identification numbers of cell culture repositories and the number of species in a genus.

Chapter 2: In this chapter, phylogenetic tree reconstruction methods are developed using COGs, GOs and pathways of whole genomes. The aim of the second chapter is to compare the phylogenetic trees developed by these noval methods with established sequence-based, gene-content-based, and ortholog-content-based trees as a means of ensuring that phylogenetic inferences are correct.

Chapter 3: The aim of the third chapter is to study the following: 1. character (genes/enzymes/pathways) evolution in a group of bacteria; 2. the evolution of genomic features (e.g. GC content, length of nucleotide sequences, and number of genes in a genome) in a group of bacterial genomes; and 3. grouping of functionally linked genes/enzymes correlated on an evolutionary level.

2. TAXONOMY ELUCIDATOR

2.1. INTRODUCTION

The taxonomy elucidator is an application designed to elucidate taxonomic nomenclature, definitions and descriptions of an organism from a stored local database for a user-selected genus name and species epithet. The Natural Language Generation (NLG) system [Jester and Porter, 1997] is followed to develop the taxonomy elucidator. It was developed using Ruby on Rails which elucidates the taxonomy of an organism in a non-ad-hoc way using the schema of the database.

The taxonomy elucidator provides comprehensive details about the taxonomy of newly sequenced microbial genomes from the GEBA project with respect to the textual model derived from *Standards in Genomic Sciences* (SIGS) genome articles [Yasawong et al., 2010, Pitluck et al., 2010]. It contains the following information: the organism's strain deposits ID, etymology, previous publications about the organism, the citations of those publications, 16S rRNA accession to the INSDC (International Nucleotide Sequence Database Collaboration) and the presence/absence details of an organism in the LPSN as sourced from the locally stored database.

2.1.1. NATURAL LANGUAGE GENERATION

Natural language generation is the task of generating texts in a specific language using key data obtained from a machine representational system like a database, tables and/or plain text. There are two major methods available for natural language generation: corpus-derived methods and template-based system. The generation of human interpretable texts from the corpus using features of corpus, linguistic rules and machine learning methods are called corpus-derived methods [Van Deemter et al., 2005]. The generation of human interpretable texts from stored templates and information retrieved from databases is called the template based NLG system [McRoy et al., 1999]. As a curated database is available for taxonomic nomenclatures and definitions, a template-based NLG system is suitable to be used to design the taxonomy elucidator given that it is relatively simple to operate when compared to the corpus-derived NLG methods. The quality of texts generated by template-based system is relatively higher than those generated by corpus-derived methods [Van Deemter et al., 2005].

2.1.2. TEMPLATE-BASED NLG SYSTEM

A template based NLG system has several stages: 1. content determination; 2. document structuring; 3.

aggregation; 4. lexicalization; 5. referring expressions; and 6. linguistic realization [Reiter and Dale, 1997]. The first three stages were adopted in order to implement the taxonomy elucidator. The latter three stages were needed for automated linguistic validations (lexicalization, referring domain element, and orthography validation). In the case of the taxonomy elucidator, the last three stages were omitted, as input data were being pre-validated for linguistic standards.

Content determination

The input for the template-based NLG system has two components, namely template and content. A template is a constant text over the output and is independent of input data provided by the user. Content is a text which varies among the generated output and is dependent on the input data. Take the following example into considerations: "The genus *Paenibacillus* currently consists of 112 validly named species". In this case, the words highlighted in red color belongs to the template and the two green elements belongs to the content. The content determination stage plays a role in defining templates and contents, where templates are drawn from published SIGS articles [Yasawong et al., 2010, Pitluck et al., 2010] and contents are obtained from the database.

Document structuring

For this stage, contents are structured together with template texts in order to construct meaningful sentences which can be interpreted by humans according to the syntax of the language. An example sentence is "The genus *Paenibacillus* currently consists of 112 validly named species". In this statement, the blue colored text is a subjective noun, the pink colored text is a verb and adverb combination and the brown colored text represents an objective noun, where contents and templates are structurally framed as "SVO" sentence pattern system of English language. The grammar and orthography of a sentence is framed in consideration of the syntax of the English language at the document structuring stage.

Aggregation

At this juncture, generated sentences are aggregated in the correct order as a means of rendering information intelligible. An example of this is the following: "... The species was first described in 1982 by Wegienek and Reddy. The species was subsequently transferred to the new genus *Actinobaculum* as *Actinobaculum suis* ...". In this statement, part of the output generated by the taxonomy elucidator is displayed, comprising two chunk statements which are connected in a specific order. The aggregation of chunks plays an important role in generating texts of readable document

quality.

The advantages of using the template-based NLG system in conjunction with the above stages are that complex sentences can be easily modulated and arranged, contents can be easily updated or improved independently, and complex mathematical strategies are not required like in corpus-derived methods for the generation of meaningful texts [Van Deemter et al., 2005].

2.1.3. ARCHITECTURE OF RAILS

Ruby on Rails is a web application framework. Ruby on Rails and other Ruby Gems (activerecord, linguistics and mapper) [Baechle and Kirchberg, 2007] were used to design the taxonomy elucidator. Gems are self-constrained format of Ruby programs and libraries.

Ruby on Rails was chosen to build the taxonomy elucidator due to its convenient MVC (Model-View-Controller) architecture [Thomas and Hansson, 2006]. The data flow is managed using efficient display logics, application control logics of MVC-structured architecture and improved software testing methods. Ruby on Rails runs on Ruby programming language that integrates controlling scripts, web server, database connector, algorithms, and Ruby libraries.

The main components in Ruby on Rails are the database, model, view, controller, web server and front end browser [Thomas and Hansson, 2006]. The software pattern for the taxonomy elucidator implementation is explained in Figure 1.

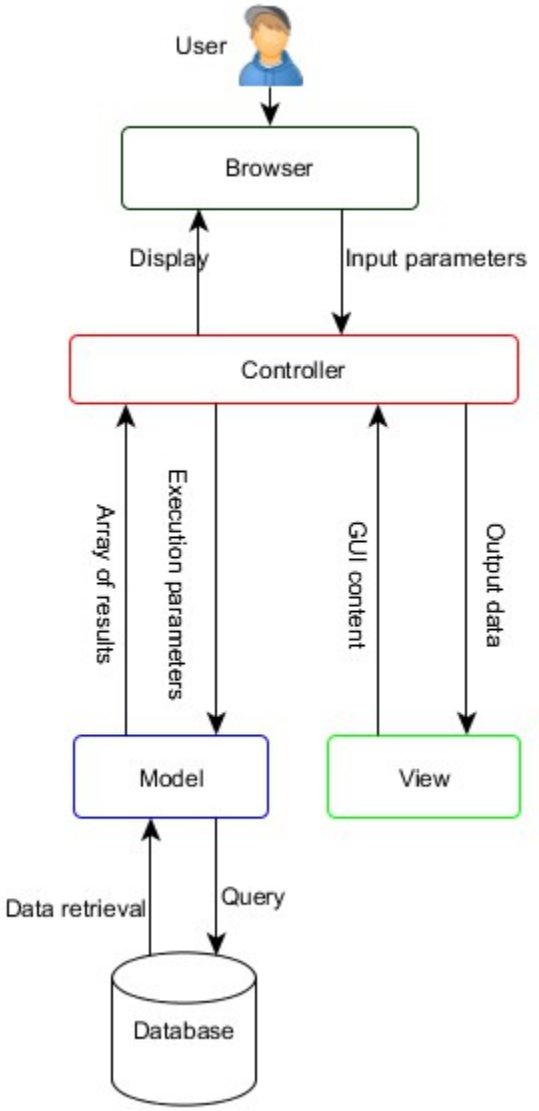


Figure 1. The MVC architecture of Ruby on Rails – a software pattern for taxonomy elucidator implementation. The view generates the GUI for output data.. The browser receives user requests and displays a GUI content from view to the user. The controller interacts with model, browser and view. The model query and retrieve data from the database.

Browser

The browser is an application for retrieving and presenting information using transfer protocols (e.g. Hyper Text Transfer Protocol). Taxonomy elucidator uses the browser to display the output and retrieve input parameters from the user.

Model

A model represents a single object or collection of structured objects in the MVC architecture [Reenskaug, 1979]. A model is a component of the MVC architecture used to correspond database and controller component. The Ruby methods for data retrieval from the database can be implemented in model using an activerecord Gem. The customization of data retrieved from the database can be accomplished by utilizing the regular expression functions in Ruby methods in model.

View

To each model, one or more views are created. A view is a visual representation of its model [Reenskaug, 1979]. A view component in Rails comprises of markup code embedded along with Ruby scripts which generate a user interface in Hyper Text Markup Language (HTML). The layout design, frame design, query retrieval system, and table design can be implemented in the view. The view can embed JavaScript style sheets that are used to customize the user interface. The view has two layouts: static background layout and dynamic layout.

Helper

The helper is a component in which the external libraries can be stored. The supportive Ruby methods for model, controller and view can be implemented in helper.

Controller

The controller is the link between the user and the system [Reenskaug, 1979]. The controller component in Rails interacts with model, view, and helper. It corresponds user request and visual output between user and other components. The main processing unit of a Rails project is the controller.

Web server

The Phusion Passenger is a web server available for Rails projects. It is an automated application which uses the Apache HTTP server. This web server system dynamically adjusts the number of processes based on server traffic. Phusion Passenger is a smart system which senses the traffic and computational time required by the application and transfers the request/response data accordingly. It stabilizes the server and prevents the application from being overburdened. Furthermore, the integration of another

Rails project within an existing Rails project is possible with Phusion Passenger.

2.1.4. DATABASE

In the taxonomy elucidator, the content data is stored in a PostgreSQL database. The relational model of data storage in a PostgreSQL database is a more suitable setup for the taxonomy elucidator. It allows the content data to evolve independently [Codd, 2001]. The database comprises 7 tables joined together using primary keys and foreign keys.

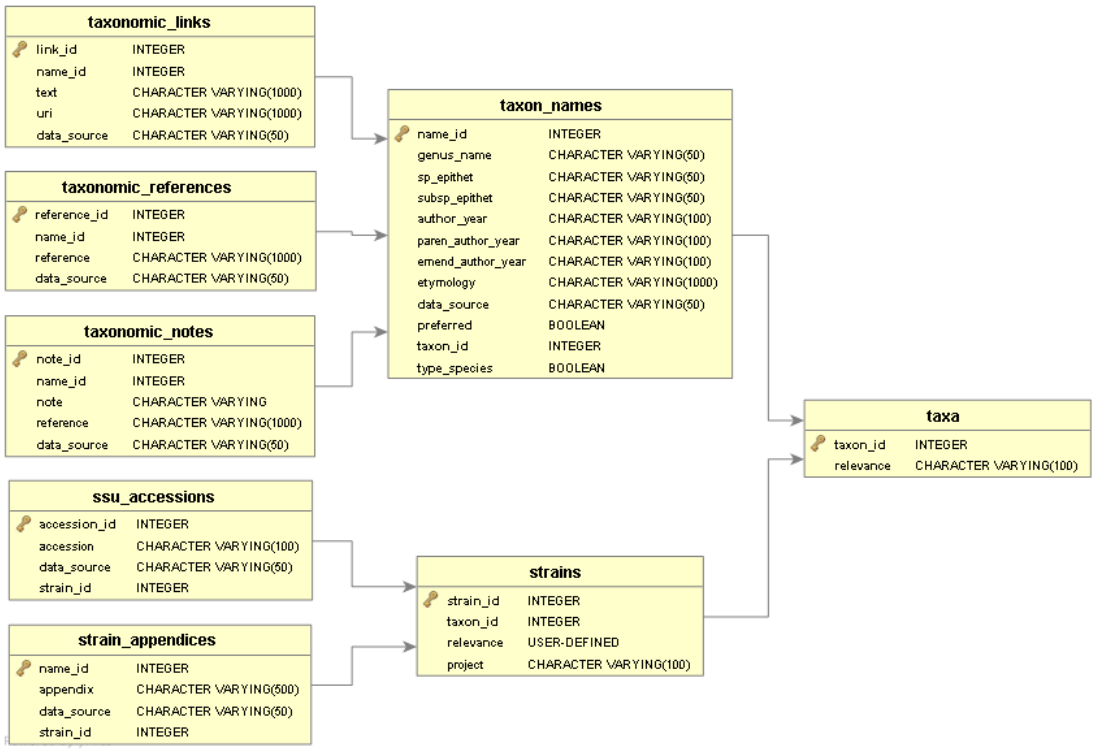


Figure 2. The schema of the database used in the taxonomy elucidator. The title of each box represents the name of the table, each row in the box represents the names of the columns in a table, the data types of the table values are represented next to the column names, and the key symbol represents the primary key of the table.

The database whose scheme is described in Figure 2 explains the content data used in the taxonomy elucidator. The database was filled with data obtained from LPSN [Euzeby 2012]. The database contains two main tables: taxon_names and strains. The taxon_id was used as a foreign key with which to join taxon_names and strains tables together. The taxon_names table includes nomenclature details

and taxonomic definitions about organisms such as genus name (genus_name), species epithet (sp_epithet), etymologies (etymology), etc. The strains table is used as an intermediate table with which to join other tables (ssu_accessions and strain_appendices) comprising additional details about strains such as 16S rRNA accession ID, hyperlink to the source of 16S rRNA, etc. The name_id in the taxon_names table was used as a primary key with which to join the tables (taxonomic_links, taxonomic_references and taxonomic_notes) comprising taxonomic descriptions such as citation for taxonomic references, notes for the references, etc.

2.1.5. STANDARDS IN GENOME SEQUENCING PROJECTS

The rapid growth of biological information requires easy access to a vast amount of data. Standard conditions were developed as a means of facilitating the discovery of biological data in an appropriate manner and the subsequent distribution of data according to the requirements of practitioners operating throughout the biological and biomedical sciences [Taylor et al., 2008]. “Minimum Information for Biological and Biomedical Information” was an initial project for developing a set of standard guidelines for sharing biological data to the wide scientific community at large. Following on from this, several consortia were formed. The aim was to accelerate the formation of mutually beneficial networks of expertise [Taylor et al., 2008]. For the genomic sciences, a consortium was formed which developed a set of standard guidelines for describing and analysing genomic data.

The increased growth of genomic data led to a proliferation of sub-genomic data such as ribosomal RNA sequences, encoded protein sequences, molecular functions in a genome, pathways in a genome, taxonomic references and other relevant information derived from genomic data. The Genomic Standards Consortium (GSC) formed in 2005 as a means of firmly integrating contextual data in a way that is readable for both humans and machines. The aim of the GSC is to promote standardized methods for describing newly sequenced genomes, the exchange of genomic data throughout the scientific community, and the subsuming of genomic data and taxonomic data of newly sequenced organisms under a single pipeline [Garrity et al., 2008].

The phylogenetic way of considering organisms for genome sequencing was initially neglected in the planing of genome sequencing projects. This led to a strongly biased representation of the recognized microbial phylogenetic diversity. To apply the systematic approach on genome sequencing projects, several Bacteria/Archaea genomes were selected solely for their phylogenetic position. The project focusing on phylogeny-driven sequencing project is called as “Genomic Encyclopedia of *Bacteria* and

Archaea" (GEBA) [Wu et al., 2009].

In conjunction with the large scale phylogeny-driven microbial sequencing projects, *Standards in Genomic Sciences* (SIGS) was created in 2008 with the support of the GSC [Garrity et al., 2009, Garrity, 2011]. The idea behind SIGS is publishing genome reports of newly sequenced microbial genomes including "Minimum Information about a (Meta) Genome Sequence (MIGS/MIMS)" standards. The scientific manuscripts published in SIGS include the interdisciplinary information such as taxonomic inferences, etymology, previous scientific history of newly sequenced genomes and genomic data [Garrity, 2011]. Since 2009, SIGS has become a primary outlet for the phylogeny-driven sequencing project GEBA [Wu et al., 2009].

The taxonomy elucidator was implemented to describe the taxonomic nomenclature and taxonomic definitions of newly sequenced GEBA project genomes.

2.2. METHODS

The taxonomy elucidator comprises three components, namely the reference model, the database and the design pattern.

The reference model was adopted from published genome reports of the GEBA project. The database has been explained in detail in the section 2.1.4. The utilization of the design pattern to elucidate taxonomic information is explained in Figure 3. The design pattern comprises the components required for data querying/retrieval, data customization, and document structuring/aggregation.

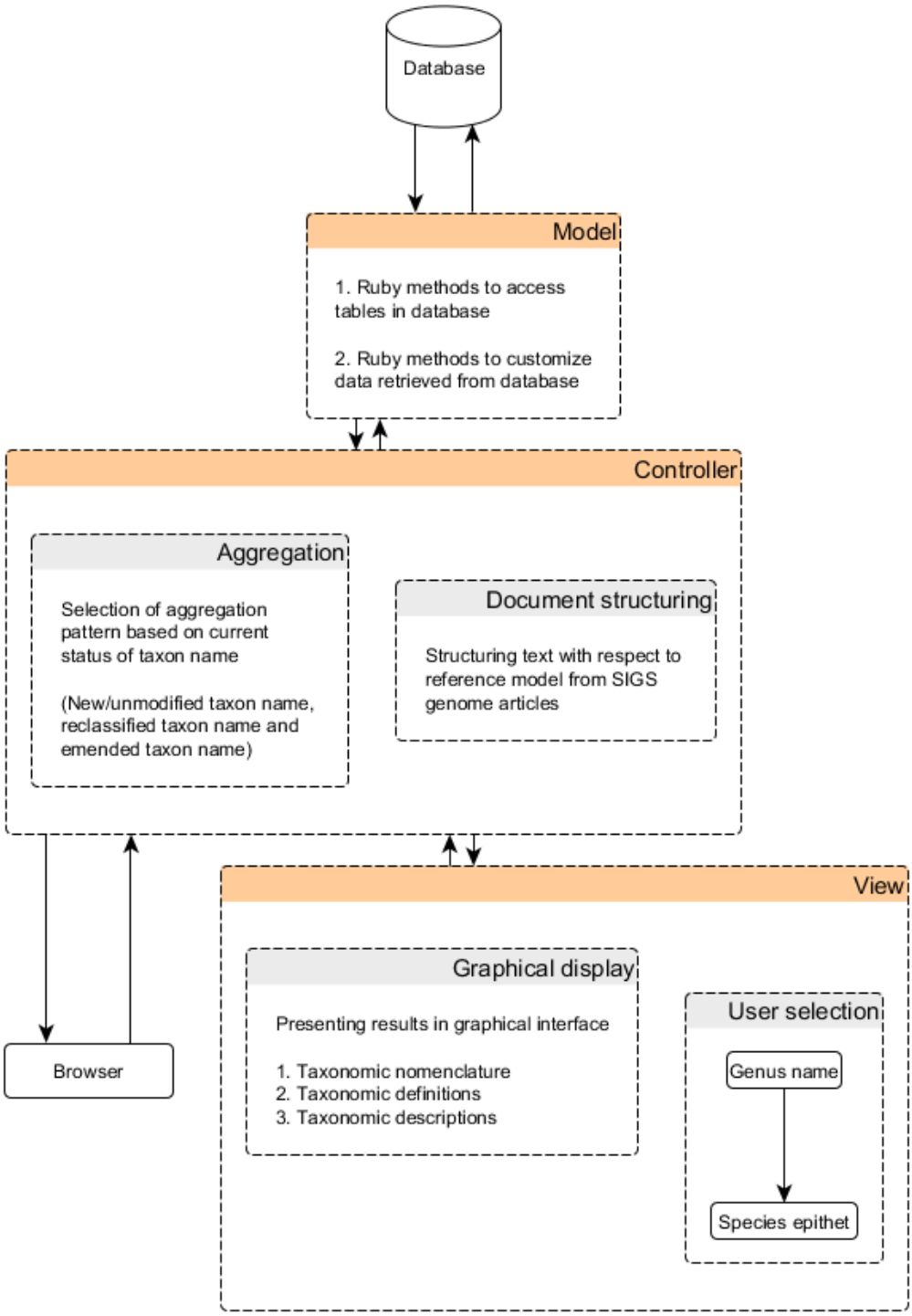


Figure 3. Design pattern of the taxonomy elucidator. The brown colored boxes represent the adopted MVC architecture components. The blue colored boxes represent the methods implemented in the taxonomy elucidator using Rails.

Data querying and retrieval

The data querying and retrieval methods were implemented in the components, model and view. The input data (genus name and species epithet) are shown as a collective list for the user to select input data. The Ruby code and activerecord Gem were used in the model to join tables in the database to query/retrieve content data (taxonomic information with respect to the query taxon name) from the database.

Data customization

The data customization methods were implemented in the model. For instance, the citation "Funke et al. 1995" is an instance of raw data retrieved from the database. The year "1995" and author names "Funke et al." are separated using regular expressions to form human interpretable textual sentences. The numbers in single digit were converted into the word equivalent (e.g. "1" into "one") using linguistics Gem.

Document structuring and aggregation

Document structuring and aggregation methods were implemented in the controller. The content data from model and template data from helper components were aggregated and structured in the controller. The Ruby methods were implemented in controller to structure the template data and content data in consideration of the reference model adopted from SIGS genome articles. The Ruby script in controller aggregates the sentences based on the current status of the taxon name of an organism.

Three aggregation patterns were followed to provide taxonomic information about an organism. If the taxon name is not modified for an organism since it was named, the aggregation pattern is called "New/unmodified taxon". If the taxon name is differently named at first and then reclassified for an organism, the aggregation pattern is called "reclassified taxon". If the taxon name is differently named at first and then emended later for an organism, the aggregation pattern is called "emended taxon".

2.3. RESULTS

User interface overview

Figure 4. Input selection screenshot - genus name and species epithet. A list of genus names and corresponding species epithets are provided as a consecutive selection for the user. It shows the selection of a genus name and the species epithet (Desulfobulbus propionicus).

Taxon name:	*Desulfobulbus propionicus* Widdel 1981
16S rRNA INSDC accession:	AY548789 [LPSN]
Etymology of the epithet:	N.L. n. acidum propionicum, propionic acid; L. masc. suff. -icus, suffix used with the sense of pertaining to; N.L. masc. adj. propionicus, pertaining to propionic acid.
Link(s):	Validation List N° 7 in IJSEM Online
Reference(s):	1. VALIDATION LIST N° 7. Int. J. Syst. Bacteriol., 1981, 31, 382-383. 2. WIDDEL (F.): Anaerober abbau von fettsäuren und benzoesäure durch neu isolierte arten sulfat-reduzier-ender bakterien. Dissertation. Georg-August-Universität zu Göttingen. Lindhorst/Schaumburg-Lippe, Göttingen, 1980.
Note(s):	
Type strain deposit(s):	1pr3 = ATCC 33891 = DSM 2032 = "Lindhorst" = VKM B-1956
Automated text:	Strain DSM 2032^T (= 1pr3 = ATCC 33891 = "Lindhorst" = VKM B-1956) is the type strain of the species *D. propionicus*, which is the type species of its genus *Desulfobulbus* [1, 2]. The genus currently consists of five validly named species [3]. The genus name is derived from the Latin word 'de-' meaning 'from', the Latin word 'sulfur' meaning 'sulfur', the Neo-Latin word 'desulfo-' meaning 'desulfuricating' and the Latin word 'bulbus' meaning 'a bulb, an onion', yielding the Neo-Latin word 'Desulfobulbus' meaning 'onion-shaped sulfate reducer' [3]. The species epithet is derived from the Neo-Latin word 'acidum propionicum' meaning 'propionic acid' and the Latin word '-icus' meaning 'suffix used with the sense of pertaining to', yielding the Neo-Latin word 'propionicus' meaning 'pertaining to propionic acid' [3]. Here we present a summary classification and a set of features for *D. propionicus* strain DSM 2032^T, together with the description of the **complete/non-contiguous finished** genome sequencing and annotation.
Text reference(s):	1. VALIDATION LIST N° 7. Int. J. Syst. Bacteriol., 1981, 31, 382-383. 2. WIDDEL (F.): Anaerober abbau von fettsäuren und benzoesäure durch neu isolierte arten sulfat-reduzier-ender bakterien. Dissertation. Georg-August-Universität zu Göttingen. Lindhorst/Schaumburg-Lippe, Göttingen, 1980. 3. Euzéby JP. List of bacterial names with standing in nomenclature: A folder available on the Internet. Int J Syst Bacteriol 1997; 47:590-592.

Figure 4 shows the input action of a user. The user interface has two input functions: genus name and species epithet name. The user can select the genus name from the list. The “Proceed” action yields the list of available species epithets for the selected genus. The combined input of the genus name and species epithet thereby initiates the taxonomy elucidator.Figure 5. Output display. It shows an example of output from the taxonomy elucidator. It contains an automatically generated table with taxonomic

definitions and scientific publication references to the selected genus name and species epithet (Desulfobulbus propionicus).

The output shown in Figure 5 contains nine rows of taxonomic information, which are listed as: 1. “Taxon name”, comprising genus name, species epithet and reference to the first publication which mentioned this taxon; 2. “16S rRNA INSDC accession”, a hyperlink to the remote INSDC server for the 16S rRNA data of the selected taxon; 3. “Etymology of epithet”, an etymological description of the species epithet; 4. “Links”, a list of available scientific publications about the taxon; 5. “References”, the literature-based evidence of published scientific articles about the taxon; 6. “Notes”, taxonomical hints about the organism obtained from LPSN; 7. “Type strain deposits”, cell culture repositories IDs from several collections for the selected taxon; 8. “Automated text”, automatically generated text which describes the taxonomic nomenclature and definitions of a selected organism with regard to the LPSN standard; and 9. “Text references”, evidence for the automatically generated text which contains taxonomic information.

Taxonomic descriptions

The different types of textual aggregation patterns were designed based on the taxonomic status of an organism. The results of those aggregation patterns are shown and discussed in this section namely: new/unmodified taxon, reclassified taxon, and emended taxon.

2.3.1. NEW/UNMODIFIED TAXON

Strain DSM 2032^{T} (= 1pr3 = ATCC 33891 = "Lindhorst" = VKM B-1956) is the type strain of the species *D. propionicus*, which is the type species of its genus *Desulfobulbus* [1, 2]. The genus currently consists of five validly named species [3]. The genus name is derived from the Latin word 'de-' meaning 'from', the Latin word 'sulfur' meaning 'sulfur', the Neo-Latin word 'desulfo-' meaning 'desulfuricating' and the Latin word 'bulbus' meaning 'a bulb, an onion', yielding the Neo-Latin word 'Desulfobulbus' meaning 'onion-shaped sulfate reducer' [3]. The species epithet is derived from the Neo-Latin word 'acidum propionicum' meaning 'propionic acid' and the Latin word '-icus' meaning 'suffix used with the sense of pertaining to', yielding the Neo-Latin word 'propionicus' meaning 'pertaining to propionic acid' [3]. Here we present a summary classification and a set of features for *D. propionicus* strain DSM 2032^{T}, together with the description of the **complete/non-contiguous finished** genome sequencing and annotation.

Figure 6. Screenshot of taxonomy elucidator produced aggregation pattern for new/unmodified taxon. The aggregation pattern comprises five sentences. An example of the user selected taxon

(Desulfobulbus propionicus) is shown in this figure.

If a species is named under a taxonomic hierarchy and the hierarchy is not modified, the taxonomic information are described in five sentences. Figure 6 shows an example of a new/unchanged taxon and its sentence aggregation pattern.

1. The first sentence provides the user selected organism's strain deposit ID of the taxon, a list of equivalent strain deposit IDs of the taxon, the genus name, and the type strain of the selected species with citations to reference publications.
2. The second sentence provides the number of validly named species under the selected genus name with citations to reference publications.
3. The third sentence elucidates the etymology of the selected genus name with citations to reference publications.
4. The fourth sentence elucidates the etymology of selected species epithet with citations to reference publications.
5. The fifth sentence provides the taxon name, strain deposit ID and constant template data.

2.3.2. RECLASSIFIED TAXON

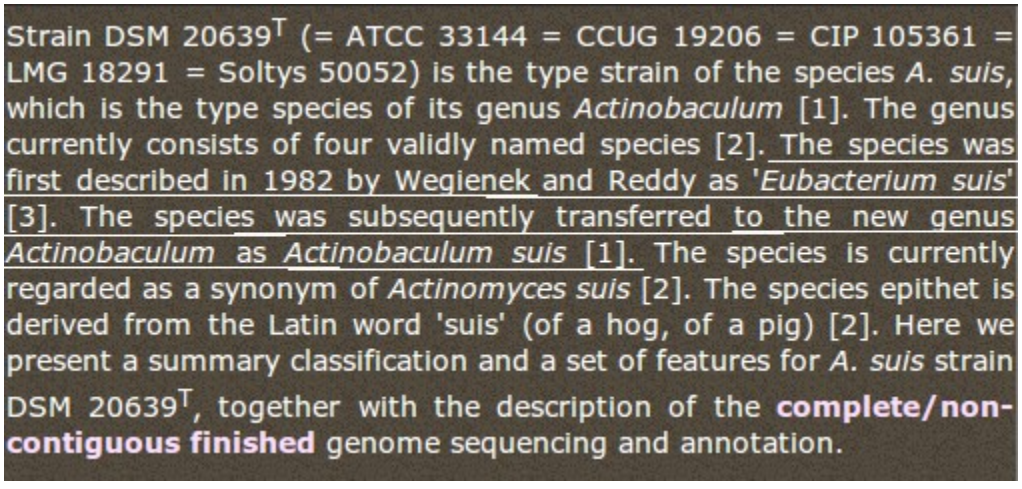


Figure 7. Screenshot of taxonomy elucidator produced aggregation pattern for reclassified taxon. The aggregation pattern comprises of seven sentences. The example of user selected taxon (Actinobaculum suis) is shown in this figure.

Figure 7 shows an example of a reclassified taxon and its sentence aggregation pattern. If an organism is named under a taxonomic hierarchy and the name is subsequently reclassified into another taxonomic hierarchy, the reclassified taxonomic information are described in seven sentences along with extra details about previous taxonomic information. *Actinobaculum suis* is an example of a reclassified taxon. In Figure 7, the taxonomic information about the *Actinobaculum suis* has two additional sentences in comparison with the previous example shown in Figure 6. These two sentences are underlined in Figure 7.

1. The first additional sentence provides the information about previous taxonomic classification of the selected taxon such as previously described year, previously described authors and genus name with citations to reference publications.
2. The second additional sentence provides the information about reclassification of the selected taxon from one genus name to an another genus name with citations to reference publications.

2.3.3. EMENDED TAXON

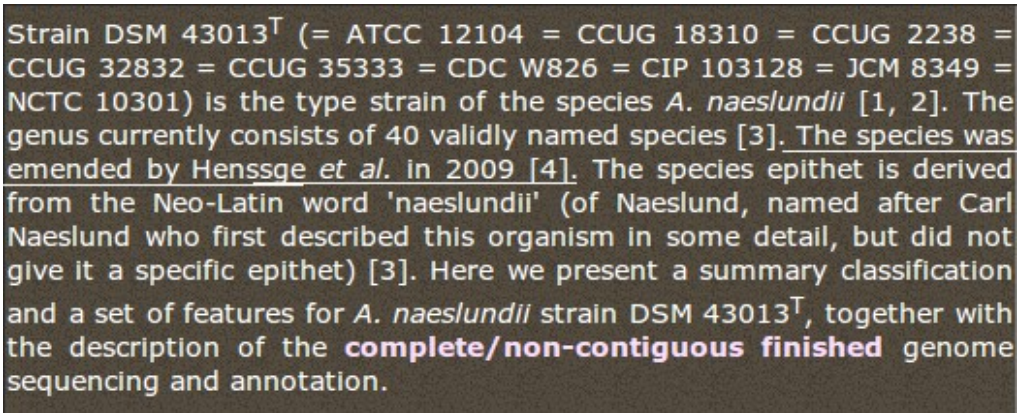

Strain DSM 43013^T (= ATCC 12104 = CCUG 18310 = CCUG 2238 = CCUG 32832 = CCUG 35333 = CDC W826 = CIP 103128 = JCM 8349 = NCTC 10301) is the type strain of the species *A. naeslundii* [1, 2]. The genus currently consists of 40 validly named species [3]. The species was emended by Henssge *et al.* in 2009 [4]. The species epithet is derived from the Neo-Latin word 'naeslundii' (of Naeslund, named after Carl Naeslund who first described this organism in some detail, but did not give it a specific epithet) [3]. Here we present a summary classification and a set of features for *A. naeslundii* strain DSM 43013^T, together with the description of the **complete/non-contiguous finished** genome sequencing and annotation.

Figure 8. Screenshot of taxonomy elucidator produced aggregation pattern for emended taxon. The aggregation pattern comprises of six sentences. The example of user selected taxon (Actinomyces naeslundii) is shown in this figure.

Figure 8 shows an example of an emended taxon and its sentence aggregation pattern. If an organism is named under a taxonomic hierarchy and the name is subsequently emended by an another author, the emended taxonomic information are described in six sentences along with extra details about previous taxonomic information. *Actinomyces naeslundii* is an example of an emended taxon. In Figure 8, the taxonomic information are provided for *Actinomyces naeslundii* with an additional sentence in

comparison with the previous example shown in Figure 6. The additional sentence is underlined. The additional sentence elucidates the emendation of the taxon with citations to reference publications.

2.4. DISCUSSION

The taxonomy elucidator was designed for 1920 genera and their corresponding species. The number of genera can be increased by updating the database. The taxonomy elucidator was validated using 20 randomly selected genera for each aggregation pattern (new/unmodified or reclassified or emended taxon) and 2-5 species including the type species of each genus. The generated texts were carefully validated for their scientific and syntactic integrity. The sentences with taxonomic nomenclature, definitions and references were generated in such a way that these can be directly apply in GEBA genome manuscripts. Some of the successfully published articles in the SIGS journal with the taxonomy elucidator generated texts include Pagani et al., 2011, Pati et al., 2011, and Ivanova et al., 2011.

The correctness of the taxonomic elucidator was validated in two stages:

Content data validation

The taxonomy elucidator generated texts were validated with respect to the data obtained manually by querying the database. The taxonomy elucidator generated data were in accordance with the manually obtained data from the database for the selected taxon. This validation was performed for random inputs as described in the previous paragraph. Overall, it resulted in clear evidence that the taxonomic information shown by taxonomy elucidator is accurate.

Document aggregation and structure validation

The congruence between the reference (document aggregation and structure) patterns adopted from previously published SIGS genome articles and the results produced by the taxonomy elucidator offers convincing evidence that the taxonomy elucidator is able to produce a text document with aggregation and structure patterns which can be used in genome manuscripts.

2.5. CONCLUSION AND FUTURE PERSPECTIVE

The taxonomy/phylogeny driven genome publications are in primary importance [Wu et al., 2009]. In conjunction with this importance, it is essential to develop an application which comprehend the taxonomic details of an organism using the taxonomic data stored in the locally established database.

As has been demonstrated. the taxonomy elucidator fulfills this requirement.

This application can be further improved for generating text with other information about the organism. The information which can be included in text generation are genome-specific data (annotation details and nucleotide/protein sequences, protein domain details), phenotype data (shape, size, motility factors and growth medium profile), data from online resources (GOLD, NCBI and KEGG), DNA preparation data (method of sequencing, source of an organism and DNA isolation method) and phylogenomic data (16S rRNA phylogenetic tree of a genus or group of organisms).

Rails has the feature of integrating several projects together. Once the local databases have been available for newly genome-sequenced organism related information, Rails projects for those information (genome-specific data, phenotype data, DNA preparation data and phylogenomic data) can be developed and integrated with the taxonomy elucidator project.

3. GENOME CLASSIFICATION USING ENCODED FUNCTIONALITIES

3.1. INTRODUCTION

The phylogenetic reconstruction using 16S rRNA for bacteria is widely accepted from the 1990s on for its conserved nature [Weisburg et al., 1991]. The availability of fully sequenced genomes ultimately results in a phylogenetic reconstruction method, which is less sensitive to the inconsistencies which arose in other methods [Delsuc et al., 2005]. The whole-genome based phylogenetic reconstruction method became the most effective strategy for classifying organisms [Snel et al., 1999], and whole-genome sequence based phylogenetic tree construction methods have been developed. "Genome BLAST distance phylogeny (GBDP)" strategy is an example for the whole-genome based method [Auch et al., 2006]. Beyond sequence-based approaches, phylogenetic reconstruction using the presence/absence or abundance of shared functionalities (e.g., pathways) and COGs are the new strategies developed in this chapter.

Three distance-based phylogeny reconstruction strategies (COG-based, pathway-based and GO-based) with 34 variations for distance calculations were formulated using shared COGs, pathways and GOs between genomes. The distances between genomes were calculated using the Sørensen's/Steinhaus similarity coefficient method and distance matrices were generated for eight sets of newly genome-sequenced organisms. From the distance matrices, phylogenetic trees were generated using the neighbor-joining method (NJ) [Saitou and Nei, 1987] especially the unweighted neighbor-joining method (UNJ) [Gascuel, 1997b]. The UNJ and BioNJ [Gascuel, 1997a] are extensions of NJ method. Both applies the same NJ selection criterion with variation in estimation and reduction formula to infer tree. The BioNJ is better than NJ for sequence data with high substitution rates [Gascuel, 1997a]. The simulation study shows that UNJ is better than NJ with respect to 10-50% of error reduction in terms of ability to recover true tree structure [Gascuel, 1997b]. Thus, UNJ was chosen as better choice among the NJ extensions for the distances calculated using shared functionalities.

The suitability of distance matrices for phylogenetic tree reconstructions was estimated using their treelikeness assessed using δ-values of all quartets of taxa [Holland et al., 2002] especially the altered δ-value called ε-value. The δ-value sums up all Q/R if $R \neq 0$ and 0 if R = 0; ε-value adds 1 if R = 0, where Q and R are sum of all q and r criteria computed from taxon quartets and defined by Guindon and Gascuel [Guindon and Gascuel, 2002]. The ε-value was chosen for its linear relation with the

compatibility-score (c-score) for test dataset and test trees (GBDP tree construction strategy) by Auch, Henz and Göker [Auch, Henz, and Göker, 2008]. The c-score is the proportion of all non-trivial splits in the query tree that are compatible with reference tree and all non-trivial splits in the query tree [Henz et al., 2005]. The c-score assess the topological accuracy of a tree with respect to the reference tree.

Phylogenetic trees already established using whole-genome data were compared to new trees with respect to Robinson-Foulds (RF) distances. The RF distance method is a method for expressing the agreement/disagreement between two phylogenetic trees containing the same taxa. The distance between trees were calculated by means of counting the total unmatched edges [Robinson and Foulds, 1981]. The c-score between trees is calculated based on the canonical decomposition theory for metrics on a finite set. The RF distance between trees is calculated based on RF metrics.

COGs were obtained through two approaches: 1. obtained from the Integrated Microbial Genomes (IMG) server [Markowitz et al., 2006] (IMG COGs); and 2. mapped from protein names using locally installed databases (mapped COGs). GOs and pathways were mapped from COGs using locally installed databases in PostgreSQL server.

3.1.1. DATASETS

In this work, dataset means a collection of several microbial genomes. In each dataset, Bacterial/Archaeal genomes were grouped based on the same genus/phylum/order/class. The datasets used in this study were named as *Planctomycetes*, *Archaeoglobi* + outgroups, *Escherichia* + *Shigella, Halobacteriales* + outgroups, *Bacteroidales, Roseobacter* clade, *Spirochaetae,* and *Actinomycetales*. Information about the organisms in each dataset is provided in Table S1 in the supplementary materials section. The annotated protein names and sequences of each genome were obtained from the genetic sequence database (Genbank) of National Institutes of Health (NIH), United States [Benson et al., 2000].

3.2. METHODS

The implementation of phylogenetic reconstruction strategies has four major steps: 1. mapping (such as protein names to COGs and COGs to pathways and GOs), 2. the generation of distance matrices and 3. tree construction using the unweighted neighbor-joining method and 4. distance matrix/tree evaluation.

The mapping system is divided into two parts: 1. annotated protein names to COGs, and 2. COGs to GOs and pathways. The first part includes lexical validation and COG mapping, while the second consists of GO terms mapping and KEGG pathways mapping on the basis of COGs.

Figure 9 shows an overview of methods applied in this chapter.

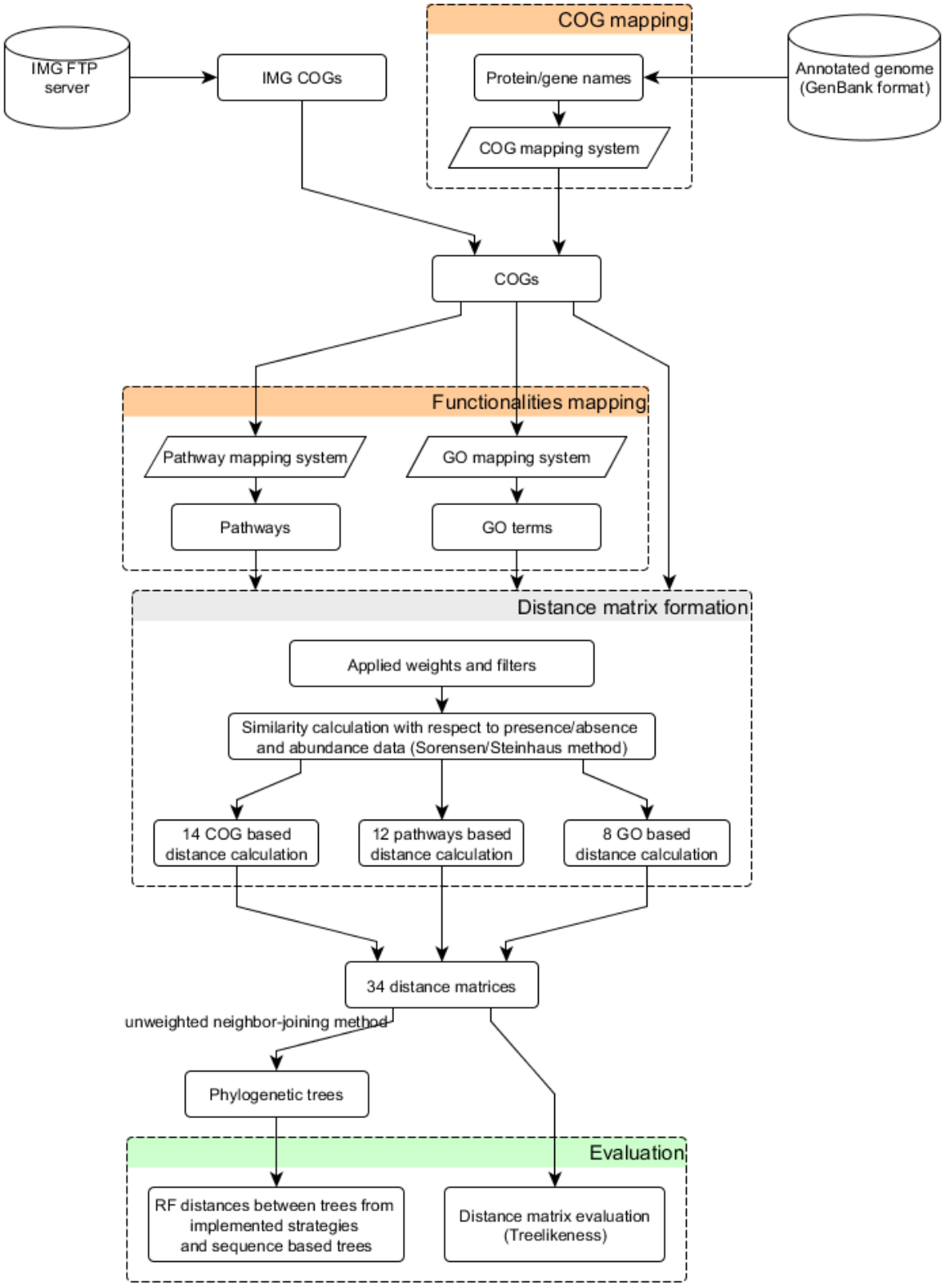


Figure 9. Phylogenetic reconstruction using functionalities. The parallelogram shapes indicate the mapping systems. The cylinders indicate the sources for mapping functionalities. The envelop cover

colors indicate the four steps followed in this work.

3.2.1. THE MAPPING SYSTEM

Naming Conventions

The correct protein/gene name with respect to the protein's correct molecular role is important. The protein/gene names themselves are highly variable, due to the fact that protein identification and naming has been carried out by diverse scientific communities. For a long time there were no naming standards available for protein/gene names. A set of naming recommendations for genes and proteins has been in place since 1957 [Miller, 1985]. Still, there are many synonymous names for individual genes/proteins. For this work, the BioLexicon was used to obtain standard gene/protein terms of annotated proteins from their synonymous names [Thompson et al., 2011].

Databases used in the mapping system

Mapping is a method which links two or more entities with regard to their key relationships. COGs were mapped from annotated protein names (mapped COGs). The GO domains (molecular functions and biological process) were mapped from COGs using the UniProtKB database [Magrane, 2011]. The KEGG pathways were mapped from COGs through KEGG cross-references [Kanehisa et al., 2010].

BioLexicon

The BioLexicon is a collection of gene terms from several existing data resources into a single, unified repository. It is enhanced by extracted gene name variants from biomedical literature. Extraction is facilitated through the inclusion of biologically appropriate verbs (e.g. “acetylise” vs “acetylize”, “co-activate” vs “coactivate”) together with information about typical patterns of grammatical and semantic behavior, which are acquired from the domain-specific text. In order to improve interoperability, the BioLexicon is modeled using the lexical markup framework, an ISO standard. The BioLexicon is available as a relational database which has a high coverage of biological entities and contains gene terms annotated with widely recognized and inter-operable unique accession numbers (e.g. UniProt Knowledge Base [UniProtKB] accession ID). The synonymous gene names are adopted from BioThesauraus [Thompson et al., 2011] in BioLexicon.

The BioThesaurus filters protein/gene names in order to remove highly ambiguous and nonsensical names [Liu et al., 2006]. Synonymous names are grouped together to avoid textual variants (e.g. “MIG-5” vs “mig-5”), punctuation (e.g. “TIMP3” vs “TIMP-3”), and syntactic variants (“tissue inhibitor of

metalloproteinase 3" vs "tissue inhibitor of metalloproteinases 3"). The multiple protein/gene names for single molecular roles indicate the relative popularity of a molecular role in scientific protein research [Liu et al., 2006].

The BioLexicon was obtained from European Bioinformatics Institute FTP website. A local copy of the BioLexicon was set up in a PostgreSQL database. Aside from reducing the ambiguity of gene names, BioLexicon provides the UniProtKB accession ID which assists the mapping of other functionalities such as GOs and pathways. It was also connected to other locally established databases (e.g. UniProtKB, GO, KEGG cross-references and KEGG relations). The size of the BioLexicon database is larger than other databases used in this study. Thus, it took more time to query and retrieve data than other databases which locally implemented and described below.

UniProtKB

The UniProtKB is a combined view consisting of a protein sequence and functional information. It has a complete collection of all publicly available protein/gene information. It consists of two sections, UniProtKB/Swiss-Prot and UniProtKB/TrEMBL. UniProtKB/Swiss-Prot comprises manually curated data. The information in each entry is annotated and reviewed by curators. UniProtKB/TrEMBL comprises automatically generated gene/protein annotations and classifications. In addition, UniProtKB has cross-references to other biological resources. The GO terms are manually curated and cross-referenced in UniProtKB in the context of Gene Ontology Annotation (UniProtKB-GOA) [Magrane, 2011].

Gene ontology

Gene ontologies are classified into three domains: molecular functions (elemental gene activities at a molecular level, e.g. catalytic or binding activities), biological processes (operations or sets of molecular events with a defined beginning and end, e.g. cell death, apoptosis and sub-processes) and cellular components (parts of a cell or an extracellular environment, e.g. nuclear inner membrane, inner envelope). Every annotation in GO is attributed to a source, which may be a literature reference or another database. The annotation indicates the type of source, thereby providing the association between the gene product and the GO term. A standard set of evidence codes is used in consideration of the different types of experimental determinations. The GO terms are normally created based on curatorial review of published literature and are supported by experimental evidence [Harris et al., 2004]. As molecular functions and biological processes state the functional characteristics of a gene,

both domains were included in phylogenetic tree reconstruction strategies. The GO database was obtained from the GO web server and locally established in a PostgreSQL database. It was also linked to the other locally established databases such as UniProtKB.

KEGG cross-references

KEGG [Kanehisa and Goto, 2000] is a web resource which provides biological information about genes, genomes and their molecular functions, orthologs, pathways, etc. KEGG has highly sophisticated databases to cross-reference their data with other biological resources. The KEGG cross-reference and KEGG relations databases were locally established. The COG-KEGG ortholog (KO) cross-reference and KO-pathway relations are some examples. The KEGG cross-reference means a database table used to map external entities of KEGG (e.g. COG). The KEGG relation means a database table used to relate internal KEGG entities (e.g. KEGG ID, KO ID, KEGG Pathway ID).

All these above databases were downloaded from remote servers and imported into the PostgreSQL database. The locally implemented databases and relations between them are shown in Figure 10.

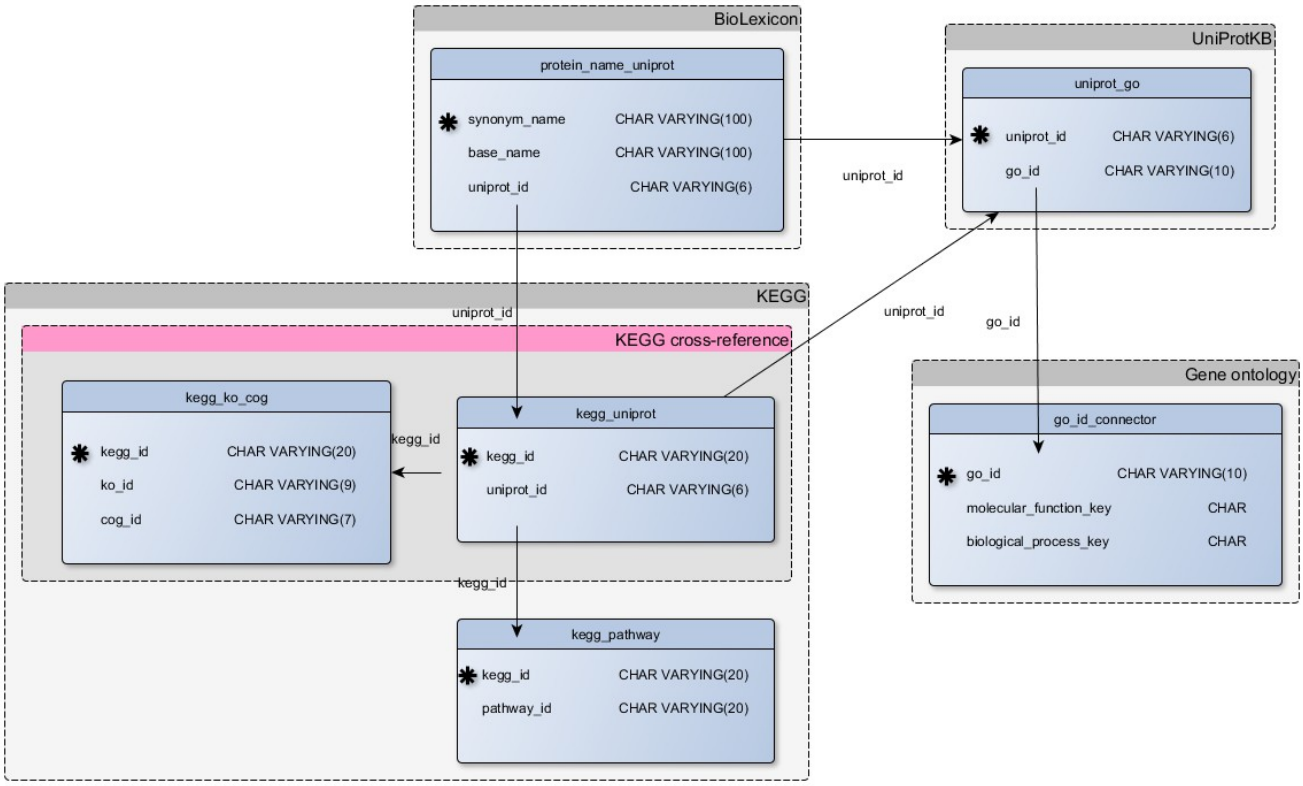


Figure 10. Schema of databases used in this chapter. The blue boxes represent tables, grey color boxes represent the databases in the PostgreSQL server, and pink boxes represent cross-reference tables. The

first column in each blue box indicates column names in the table and the second column indicates the data type. The arrow indicates the relation established between tables. The black start represents the primary key of the table.

Protein name to COG mapping

The annotated protein names from genomes were lexically validated and mapped to COGs.

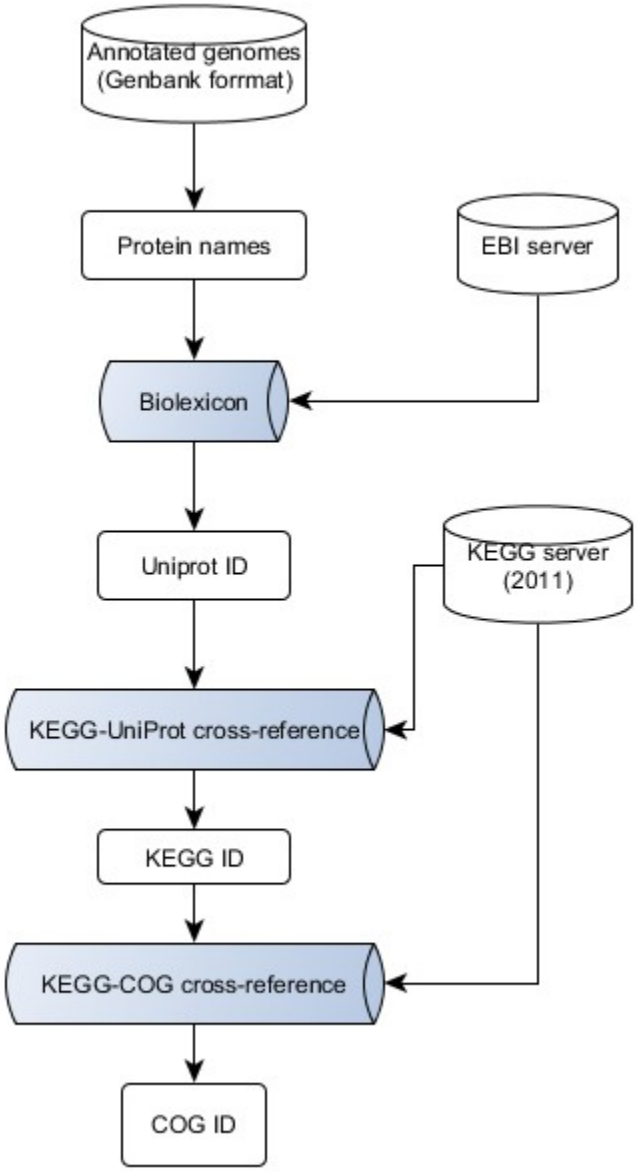


Figure 11. Annotated protein name to COG map. The map explains the serial connections established to obtain COG from the protein name. The vertical cylinder structure indicates the source of each

database/database table. The horizontal cylinder indicates the database/table.

A PostgreSQL database table containing the synonym name, base name and UniProt ID were locally created using the downloaded BioLexicon from the EBI server. The annotated protein names were mapped to corresponding UniProt ID using the locally created database table. The COGs were mapped from UniPro ID using locally stored KEGG cross-reference tables downloaded from KEGG server in 2011. Figure 11 shows the connections established between databases to map COGs from annotated protein names.

Mapping of COG to GO

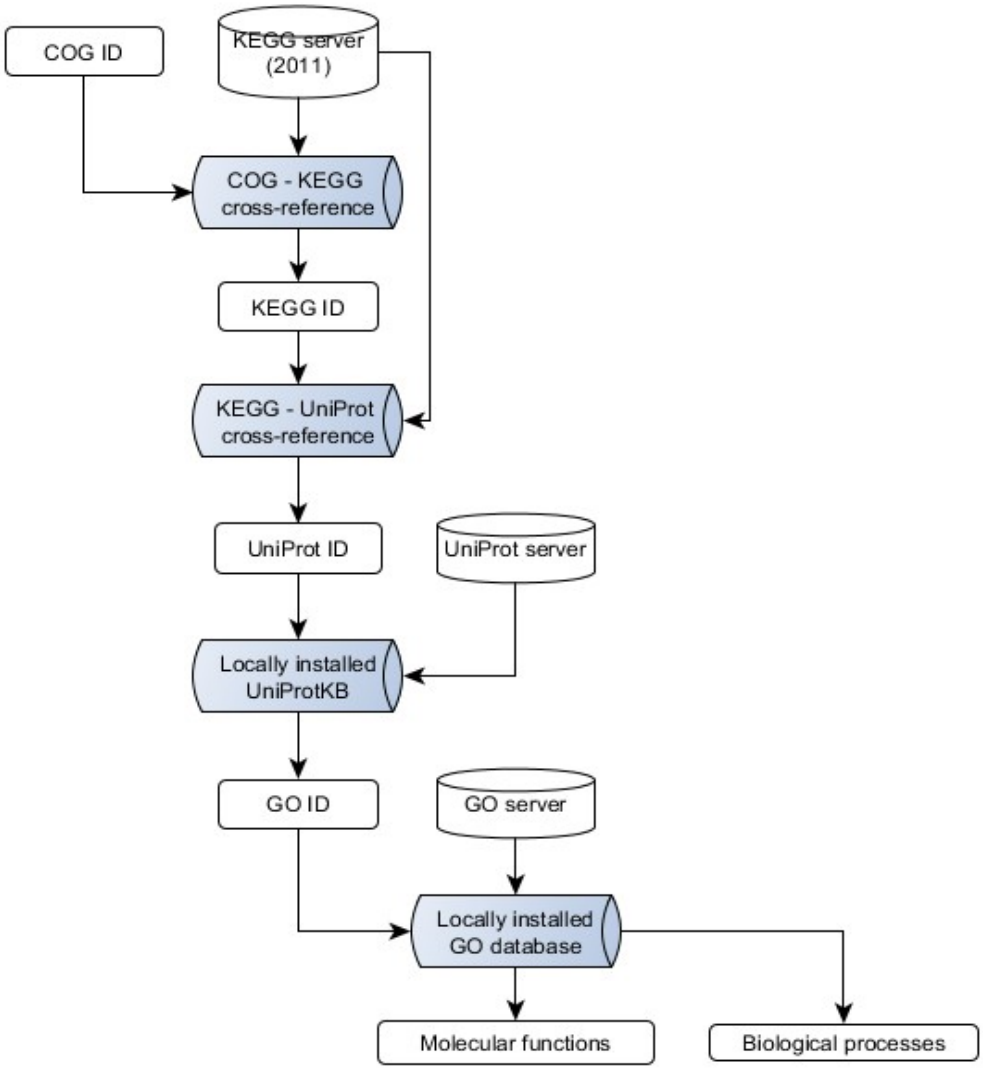


Figure 12. COG to GO domain terms map. The map explains the serial connections established to

obtain GO domain terms. The vertical cylinders indicate the source of the locally installed databases. The horizontal cylinders indicate the locally installed databases.

The COG IDs were mapped to the corresponding KEGG IDs using the COG – KEGG cross-reference database table. The KEGG IDs were mapped to the corresponding GO terms using the KEGG – UniProtKB cross-reference table and UniProtKB database. The GO terms were classified under two domains (molecular functions and biological process) using the table available in the Gene Ontology database. Figure 12 shows the connections established between databases to map GO domain terms from COGs.

Mapping of COG to pathway

The COG to pathway map does not need relations between two different databases. The relations were made using KEGG relations and cross-references.

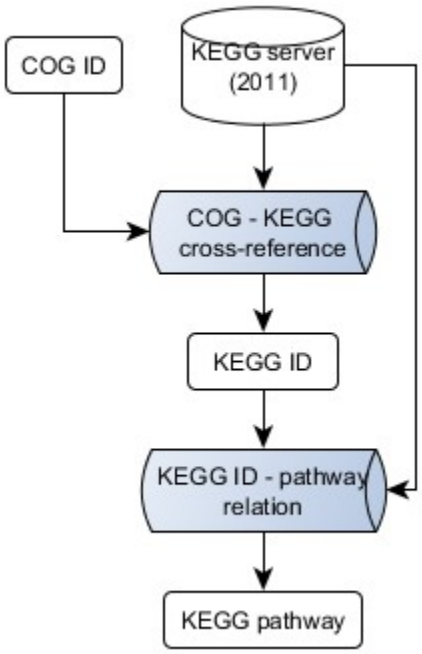


Figure 13. COG to pathway map. The map explains the serial connections established to obtain pathway. The vertical cylinder indicates the source of the locally installed database. The horizontal cylinder indicates the locally installed database tables.

The pathways were mapped from COGs using the COG-KEGG cross-reference and KEGG relations

between KEGG ID and pathway. Figure 13 shows the connections established between database tables to map pathways from COGs.

The whole mapping system was implemented and automated using the Ruby programming language and PostgreSQL.

3.2.2. SIMILARITY CALCULATION

A distance matrix is a set of distances between variables represented in the form of a matrix. The variables are arranged in a matrix. In this work, pairwise distances between genomes were calculated from similarities obtained with respect to presence/absence or an abundance of functionalities in a group of genomes. The distance matrices were used to create the distance-based phylogenetic trees.

As each COG can include several proteins/genes in a genome, the same COG can be mapped several times in a genome for different proteins. In GO terms, multiple occurrences of the same molecular function and biological process from different proteins were frequently observed. Thus, both presence/absence and abundance approaches were included to calculate the distances using COGs and GOs. The presence/absence approach was only considered to calculate the distance using pathways.

The similarity between genomes was calculated using set theory concepts. The similarities were calculated based on asymmetric similarity coefficient calculation methods. Jaccard's coefficient and Sørensen's coefficient are simple and traditional methods generally used for the calculation of asymmetric similarity coefficients [Legendre and Legendre, 1998]. Each genome was considered as a set of functionalities.

Jaccard's similarity coefficient

For two sets, $Similarity\,coefficient = \frac{a}{(a+b+c)}$ where a = a number of shared elements between two sets (S_1 and S_2), b = a number of elements present only in the first set, and c = a number of elements present only in the second set. The Jaccard distance between two sets can be calculated from the Jaccard's similarity coefficient by subtracting the similarity coefficient value from the value 1.0 [Legendre and Legendre, 1998].

Sørensen's similarity coefficient

For two sets, $Similarity\,coefficient = \frac{2a}{(2a+b+c)}$ where a, b, and c represent the same values as the

Jaccard's coefficient. Sørensen's coefficient method assigns double weight to shared elements. The Sørensen distance between two sets was calculated from the Sørensen similarity coefficient by subtracting the similarity coefficient value from the value 1.0. In the case of similarity calculation with abundance data, the Steinhaus similarity coefficient calculation method was used. It is an extended method of the Sørensen method for abundance data [Legendre and Legendre, 1998].

Steinhaus similarity coefficient

For two sets of abundance data, the similarity coefficient calculation formula is same as Sørensen's similarity coefficient where a = $Min(a_{S1}, a_{S2})$, b and c are same as the Sørensen's coefficient.

Sørensen's/Steinhaus method was selected in order to implement phylogenetic reconstruction strategies, as the Jaccard's method can not be used for abundance data.

3.2.3. APPLIED WEIGHTS

COG weights

In the case of mapped COGs, a single protein name was mapped to several COGs due to the lexical ambiguity. Even though the BioLexicon reduces much of the ambiguity in the identification process of standard protein names, the ambiguity was not completely resolved. For one protein, several COGs were mapped ($p:=(c_1 \ldots c_n)$ where p = protein/gene, c = COGs, n = number of COGs mapped per protein/gene). In this case, weighting schemes were required. For instance, thioredoxin reductase was mapped to three COGs: COG0492, COG0526 and COG0450. Each ambiguous map of an annotated protein name was weighted with respect to the proportion of ambiguity ($p:=(w(c_1) \ldots w(c_n))$ where w = weight (1/n), p = protein/gene, c = COG). From the above example, the weights for those COGs were assigned as (0.33 (COG0492), 0.33 (COG0526), 0.33 (COG0450)).

COGs weights

A set of COGs (G) in each genome was used used for similarity calculation ($c_1 \ldots c_g$, where c=COG, g = number of COGs in the genome G). The similarity calculation method using the presence/absence of COGs takes the highly weighted COG into considerations when dealing with the same COG with different weights ($m(c_x)=Max(w_1(c_x) \ldots w_n(c_x))$, where $m(c_x)$ = maximum weight assigned in the mapping system for the COG X in a specific genome). Take the following example in consideration: "(thioredoxin reductase = 0.33 (COG0492), 0.33 (COG0526), 0.33 (COG0450)), (periplasmic protein

thioldisulphide oxidoreductase DsbE = 0.5 (COG0526), 0.5 (COG0452)) and (Thiol-disulfide isomerase = 1.0 (COG0526))". From the above example, COG0526 was weighted for three different cases in a genome with different weights (0.33 (COG0526), 0.5 (COG0526) and 1.0 (COG0526)). The maximum weight 1.0 (COG0526) was considered for calculating the similarity coefficient using the presence/absence approach.

GO

The two GO domains (molecular functions and biological processes) were used in the presence/absence and abundance similarity calculation without weights. Together, two strategy variations for the presence/absence of GO domains and two strategy variations for the abundance of GO domains were formulated.

Pathway weights

The pathways were mapped from COGs. Several proteins play a role in each pathway. But, all proteins encode the pathway may not be annotated in a genome. It means that weighing schemes were necessary for calculating similarities. Take the following example in consideration: "the flagellar assembly pathway employs 41 proteins". However, 24 proteins may be annotated in a genome (G) and not remaining proteins. Thus, weights (w) were applied for each pathway as a ratio of the number of proteins annotated in a genome X per pathway to the total number of proteins described in the pathway.

Presence/absence of data based similarity calculation

The variables a, b, and c described below are the elements used in Sørensen's similarity calculation formula for presence/absence of COGs, GOs, and pathways based strategic variations.

$a = \sum_{x=1}^{g_s} (G_1(m(c_x)) \times G_2(m(c_x)))$ where c = COG/GO/pathway, G_1 & G_2 = two genomes, g_s = number of shared COGs/GOs/pathways between two genomes, m = maximum weight applied in COG/pathway mapping system for the specific COG/pathway. In the case of GO, m is always 1.0.

$b = \sum_{x=1}^{g_1} G_1(m(c_x))$ where g_1 = number of COGs/GOs/pathways present only in first genome (G_1).

$c = \sum_{x=1}^{g_2} G_2(m(c_x))$ where g_2 = number of COGs/GOs/pathways present only in second genome (G_2).

Abundance data based similarity calculation

In the case of abundant COGs, the sum of all weights was used as the weight for the similarity calculation ($t(c_x)=\sum(w_1(c_x)...w_n(c_x))$ where t(c_x) = sum of all weights assigned in the mapping system for the COG X in a specific genome). From the same example above, weights 0.33, 0.5, 1.0 assigned to COG0526 were added together as 1.83 and it was considered in order to carry out the abundance approach.

The variables a, b, and c described below are the elements used in Steinhaus similarity calculation formula for abundance of COGs and GOs based strategic variations.

$a=\sum_{x=1}^{g_s}[Min(G_1(t(c_x)),G_2(t(c_x)))]$ where c = COG/GO, G_1 & G_2 = two genomes, g_s = number of shared COGs/GOs between two genomes, t = sum of all weights applied in COG mapping system for the specific COG. In the case of GO, m is the number of abundance for each GO.

$b=\sum_{x=1}^{g_1}G_1(t(c_x))$ where g_1 = number of COGs/GOs present only in first genome (G_1).

$c=\sum_{x=1}^{g_2}G_2(t(c_x))$ where g_1 = number of COGs/GOs present only in first genome (G_1).

After applying weights to each COG/pathway, several filter schemes were applied before calculating the similarities between genomes. The applied filter schemes were: no weights, no filter, above 0.25, above 0.5, above 0.75 and 1.0. The filtered COGs/pathways were used in similarity calculations and distance matrices were developed for each filter scheme.

The unambiguous IMG COGs were downloaded from IMG server for each genome and used directly in presence/absence and abundance similarity calculation without weights.

3.2.4. WEIGHTS AND FILTERS

Weights and filters were applied to the mapped COGs. Six strategy variants were formulated using the presence/absence of COGs and six more for using the abundance data. Two strategies were created for the presence/absence and abundance of IMG COGs. In total, 14 variations were formulated for COGs. In Table 1, strategy variations formulated using COGs were collectively described. After filtering, COGs were used for similarity calculations with the assigned weights without scaling up to complete presence in all filter schemes.

Table 1. The classification of cells from first to last columns in this table explains the method of creating strategic variation using COGs with respect to mapping source, similarity calculation methods and filtered weights for similarity calculations.

Mapping source	**Data observation and similarity calculation method**	**Weights and filters**
Annotated protein names	Presence/absence of COGs and Sørensen's similarity calculation	Unweighted Weighted (0.0 – 1.0) Weighted > 0.25 Weighted > 0.5 Weighted > 0.75 Weighted (1.0)
	Abundance of COGs and Steinhaus similarity calculation	Unweighted Weighted (0.0 – 1.0) Weighted > 0.25 Weighted > 0.5 Weighted > 0.75 Weighted (1.0)
IMG COGs	Presence/absence of COGs and Sørensen's similarity calculation	Weighted (1.0)
	Abundance of COGs and Steinhaus similarity calculation	Weighted (1.0)

Weights and filters were applied to the mapped GO domain terms from annotated protein names and IMG COGs. The Sørensen's and Steinhaus similarity calculations methods were applied for presence/absence and abundance of GO domain terms. Four strategy variations for the GO terms mapped from annotated protein names and four more variations for the GO terms mapped from IMG COGs were formulated. In total, eight variations of the strategy were formulated. In Table 2, strategy variations formulated using GO were collectively described.

Table 2. The classification of cells from first to last columns in this table explains the method of creating strategic variation using GOs with respect to mapping source, similarity calculation methods and filtered weights for similarity calculations.

Mapping source	**GO domain**	**Data observation and similarity calculation method**
Annotated protein names	GO molecular functions	Presence absence GO terms and Sørensen's similarity calculation
		Abundance of GO terms and Steinhaus similarity calculation
	GO biological process	Presence absence GO terms and Sørensen's similarity calculation
		Abundance of GO terms and Steinhaus similarity calculation
IMG COGs	GO molecular functions	Presence absence GO terms and Sørensen's similarity calculation
		Abundance of GO terms and Steinhaus similarity calculation
	GO biological process	Presence absence GO terms and Sørensen's similarity calculation
		Abundance of GO terms and Steinhaus similarity calculation

Weights and filters were applied to the mapped pathways. Six strategy veriations were formulated using the presence/absence of pathways mapped from IMG COGs and six more using the presence/absence of pathways obtained from mapped COGs. In total, 12 variations were formulated for pathways. In Table 3, strategies formulated using pathways were collectively described. After filtering, pathways were used for similarity calculation with the assigned weights without scaling up to complete presence.

Table 3. The classification of cells from first to last columns in this table explains the method of creating strategic variation using pathways with respect to mapping source, similarity calculation methods and filtered weights for similarity calculations.

Mapping source	**Data observation and similarity calculation method**	**Weights and filters**
Annotated protein names	Presence/absence of KEGG pathways and Sørensen's similarity calculation	Unweighted Weighted (0.0 – 1.0) Weighted > 0.25 Weighted > 0.5 Weighted > 0.75 Weighted (1.0)
IMG COGs	Presence/absence of KEGG pathways and Sørensen's similarity calculation	Unweighted Weighted (0.0 – 1.0) Weighted > 0.25 Weighted > 0.5 Weighted > 0.75 Weighted (1.0)

The similarities were converted into distances and distance matrices were formed. A computational pipeline was set up to generate distance matrices for allstrategies mentioned in Tables 1-3. In all, 34 variations of the three strategies were implemented.

3.2.5. DISTANCE MATRIX EVALUATION

The suitability of distance matrices for phylogenetic tree construction was verified by assessing the treelikeness for each distance matrix [Holland et al., 2002]. The treelikeness was assessed using ε-values computed from quartets of taxa [Auch, Henz, and Göker, 2008]. To find a suitable distance matrix to reconstruct a phylogenetic tree, a regression model was created using the ε-values of each distance matrix as a dependent variable. 34 varitions and eight datasets were used as explanatory variables.

Multiple linear regression models

Multiple linear regression is a statistical method which is used to model a relationship between a dependent variable and more than one explanatory variable [Rubinfeld, 2000]. The strategic variations applied to calculate similarity were considered as explanatory variables. The data observation approach (presence/absence or abundance), dataset names (e.g. *Roseobacter* clade), various filter schemes applied on assigned weights and tree reconstruction strategies (e.g. COG-based) were explanatory variables. The stepwise regression was performed to remove insignificant explanatory variables. Both, forward selection and backward elimination were performed. The forward selection approach adds variables one by one from null variable with respect to low p-value until the model improves as ultimate one with the maximum possible number of variables.. The backward elimination approach removes variables one by one from all variables with respect to high p-value until the model improves as ultimate one with the possible number of variables. The threshold (0.05) was applied to p-values obtained from multiple linear regression model for each explanatory variable after stepwise removal [Faraway, 2002].

3.2.6. TREE CONSTRUCTION

The distance matrices were created in EPF (extended PHYLIP format). The EPF format is used as tree reconstruction software FastME [Desper and Gascuel, 2002] accept EPF format. The same format is used as an input for several tree construction softwares. The first row in EPF files provides information about the number of taxa. The first column provides the taxa indices or taxon names.

Unrooted trees were created using the software FastME by selecting the option “unweighted neighbor-joining method”.

3.2.7. DISTANCE BETWEEN TREES

The calculation of RF distances between sequence-based phylogenetic trees and functionality based trees provides considerable new details into the degree of topological agreement/disagreement between them.

The sequence-based trees constructed using the concatenated alignment of genomic and gene sequences were used as reference trees to study the congruence with the newly reconstructed trees. Figure 14 explains the tree reconstruction strategy using the concatenated alignment of sequences [Spring et al., 2010]. These reference trees were obtained from the pipeline implemented by my

colleague Carmen Scheuner.

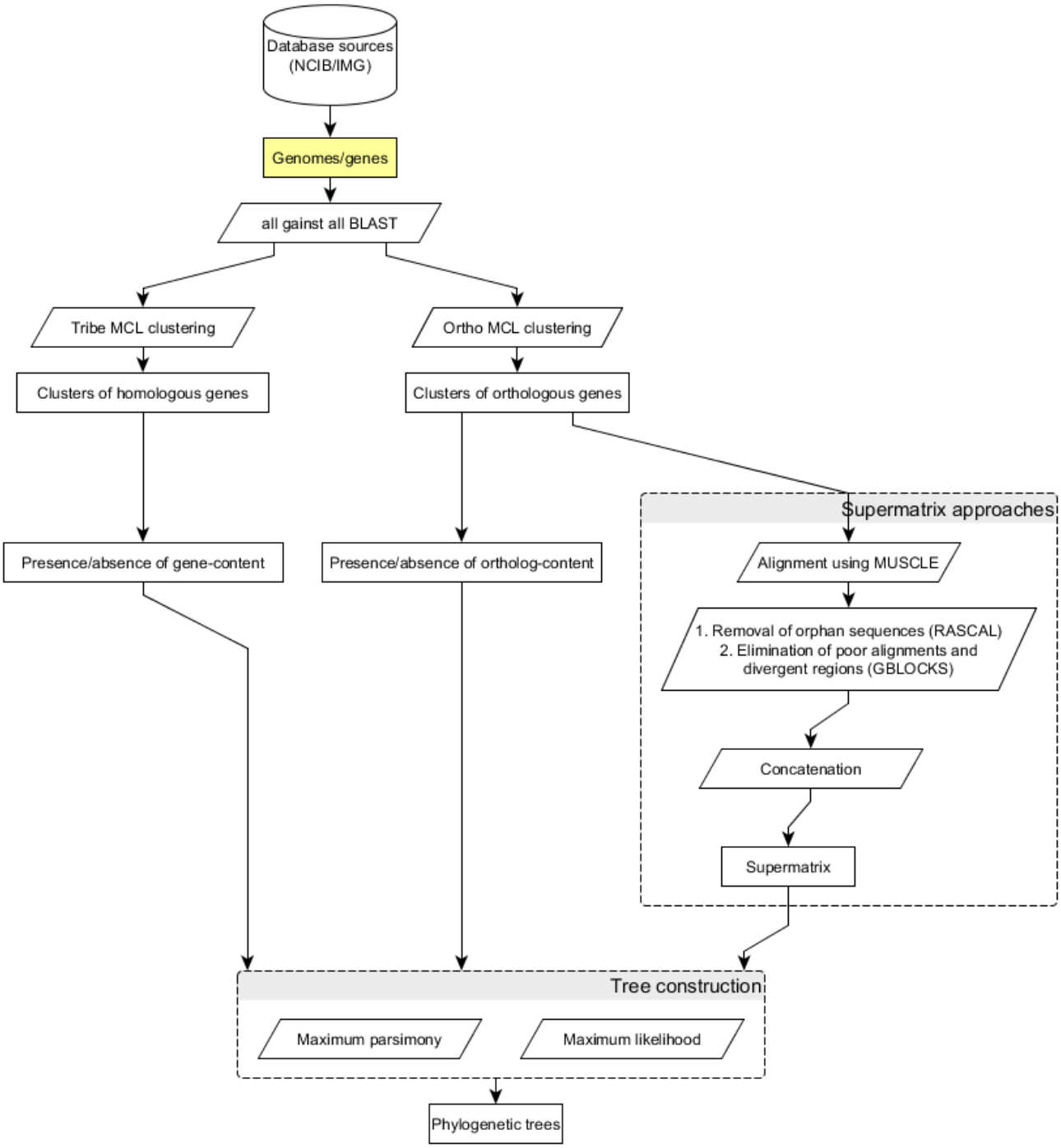


Figure 14. The method followed by Carmen Scheuner to create reference phylogenetic trees. The parallelogram indicates the method, square indicates the data, and yellow colored box indicates the input data.

Several reference trees were created using genomic/gene sequences with applied gene filters on supermatrix approaches by my colleague Carmen Scheuner. The gene filter approaches were core-genes [Medini et al., 2005], Ciccarelli filter [Ciccarelli et al., 2006], MARE (Matrix Reduction) filter [Meusemann et al., 2010], Noisy filter [Dress et al., 2008], MARE-Noisy filter (combination of both filters) and Wu & Eisen filter [Wu and Eisen, 2008].

Core-genes filter

The genes present in all genomes of the dataset were considered as core-genes. A bacterial genome can be described by its pan-genome. It composed of core genome (genes present in all genome of dataset), dispensible genome (genes present in one or more genomes in the dataset) and unique genes [Medini et al., 2005]. The core genome was considered to reconstruct phylogeny in this approach.

Ciccarelli filter

Ciccarelli and colleagues proposed to reconstruct phylogenies with the filtered 31 universally occurring genes which are more informative to reconstruct phylogeny. The genes were obtained by systematic detection and exclusion of genes that could be gained through horizontal gene transfer [Ciccarelli et al., 2006].

MARE filter

The number of genes were reduced with regard to the heuristic reduction approach. MARE software filters genes with regard to the informativeness of a single gene [Meusemann et al., 2010].

Noisy filter

The sequences were filtered using the Noisy software which removes the phylogenetically uninformative homoplastic sites and it increases the informativeness of the phylogeny [Dress et al., 2008].

Wu & Eisen filter

The filter was applied based on the potentially informative 31 protein markers suggested by Wu and Eisen. Most of those markers are housekeeping genes [Wu and Eisen, 2008].

Including the above trees, the trees created based on the gene-content in whole genome, ortholog-content in whole genome and full sequences by Carmen Scheuner were used as reference trees. Among those reference trees, gene-content based maximum-likelihood tree of *Spirochaetae* dataset was

published [Abt et al., 2012 and Anderson, 2011]. The trees were created using maximum likelihood and maximum parsimony methods. All these trees were adopted and the RF distances between reference trees and functionality based trees were then computed accordingly.

The multiple regression model was developed using RF distances between trees as a dependent variable and explanatory variables: 1. the approaches applied on sequence-based methods (e.g. Ciccarelli, MARE), 2. tree construction methods (e.g. maximum likelihood, maximum parsimony), 3. datasets (described in section 3.1.5), 4. various functionality based methods and 5. weighting and filter schemes applied in functionality based methods.

3.2.8. PRINICIPAL COORDINATE ANALYSIS

Principal coordinate analysis (PCoA), a multidimensional scaling method, was used to visualize the continuity of agreements and disagreements between phylogenetic trees based on RF distances between them. According to this method, the computed eigenvectors from the RF distances between trees were ranked from highest to lowest. The two highest vectors were plotted and discussed in the results.

3.3. RESULTS

3.3.1. DISTANCE MATRIX EVALUATION

Two regression models were created using ε-values of each distance matrix as dependent variable. The first one was created using weighted and unweighted strategies as explanatory variables, while the second one utilized filters applied weights and unweighted strategies as explanatory variables. The two different models were created to understand the impact of applied weights on functionalities over ε-values of distance matrices.

Multiple regression model – weighted and unweighted functionalities based strategic variations

This model includes the distances matrices generated using mapped COGs, GO and pathways mapped from protein names with applied weights and unweighted. Both types (presence/absence and abundance) of similarity calculation strategies were included. Three explanatory variables were used against the dependent variable (ε-values of distance matrices). The multiple regression analysis with stepwise (both forward & backward directional) removal of insignificant explanatory variables were conducted using R programming language. The explanatory variables considered in this model are shown below.

1. Explanatory variable 1: datasets mentioned in section 3.1.1
2. Explanatory variable 2: presence/absence and abundance of COGs, GO biological processes, GO molecular functions and pathways obtained from protein names.
3. Explanatory variable 3: weighted (without applied filters) and unweighted functionalities.

After stepwise removal, threshold of 0.05 was applied to p-values. Several explanatory variables were filtered out using the stepwise removal process and threshold filter. The only significant explanatory variable was “presence/absence of COGs” with the p-value of 0.00114 and estimated regression coefficient of -0.09792.

Multiple regression model – filters applied weights and unweighted functionalities based strategic variation

The model includes same distance matrices used in the previous model, distance matrices generated using applied filters (above 0.25, above 0.5, above 0.75 and 1.0) on weights for mapped COGs and pathways from protein names, and distance matrices generated using IMG COGs mapped

functionalities as a source. The selection of explanatory variables, dependent variables and regression analysis procedure were same as above model. The explanatory variables considered in this model are shown below.

1. Explanatory variable 1: datasets mentioned in section 3.1.1
2. Explanatory variable 2: presence/absence and abundance of COGs, GO biological process, GO molecular functions and pathways obtained from protein names and IMG.
3. Explanatory variable 3: weighted (above 0.25, above 0.5, above 0.75 and 1.0) and unweighted functionalities.

After stepwise removal, threshold of 0.05 was applied to p-values. Several explanatory variables were filtered out using the stepwise removal process and threshold filter. The significant explanatory variables were "presence/absence of COGs" and "weighted functionalities with filter value 1.0". The corresponding p-values were 0.002581, and 0.010606. The estimated regression coefficients were -0.0979166, and -0.0826031.

From the observations of the previous two models, the explanatory variable "presence/absence of COGs" was observed as significant explanatory variable. From the observations of the second model, one more explanatory variable "weighted functionalities with the filter 1.0" was observed as significant explanatory variable. From both models, the significant explanatory variables were observed with negative regression coefficient.

From the above observations, the distance matrices generated using the presence/absence of COGs (explanatory variable) were showing decreasing trend (negative regression coefficient) of ε-values (dependent variable) especially with the applied filter 1.0 on weights. It means that treelikeness of presence/absence of COGs based distance matrices were increasing in trend.

3.3.2. DISTANCE BETWEEN TREES

The multiple regression model was created for the reference and reconstructed trees of 34 strategic variations with the RF distance between trees as dependent variable and five explanatory variables. The regression analysis procedure was same as above two models in the section 3.3.1.

1. Explanatory variable 1: datasets mentioned in section 3.1.1
2. Explanatory variable 2: Ciccarelli filters, MARE filters, Noisy filters, MARE-Noisy filters, full

sequence, Wu & Eisen filters, core-genes, gene-content, and ortholog-content.

3. Explanatory variable 3: maximum likelihood and maximum parsimony methods.
4. Explanatory variable 4: strategic variations using functionalities mapped from IMG COGs and protein names.
5. Explanatory variable 5: presence/absence and abundance of functionalities with weights, without weights and applied filters on weights.

After stepwise removal, threshold of 0.05 was applied to p-values. Several explanatory variables were filtered out using the stepwise removal process and threshold filter. The significant explanatory variables were “IMG COGs based approach”, “GOs mapped from IMG COGs based approach”, “gene-content”, “ortholog-content”, “MARE filtered” and “presence/absence strategy”. The corresponding p-values were 10^{-16}, 10^{-16}, 10^{-12}, 0.001673, 0.0045, 0.007126. The estimated regression coefficients were -0.279533, -0.109257, -0.080688, -0.029333, -0.02646, -0.019744.

From the observations of the model, the trees created using those (significant explanatory variables from regression model) approaches were showing decreasing RF distances between trees (dependent variable) in trend with respect to negative regression coefficient. It means that the trend was increasing for the topological similarity between those trees.

3.3.3. COMPARATIVE ANALYSIS

The two highest eigenvalues from the PCoA analysis were plotted as a XY graph and it is shown in Figures 15 and 16 for eight datasets. Each tree reconstruction method for each dataset had its own XY value. All of the XY values were plotted on the XY plane, and the XY values of the tree construction methods which were close to each other on the XY plane can be considered as topologically similar trees with regard to RF distance.

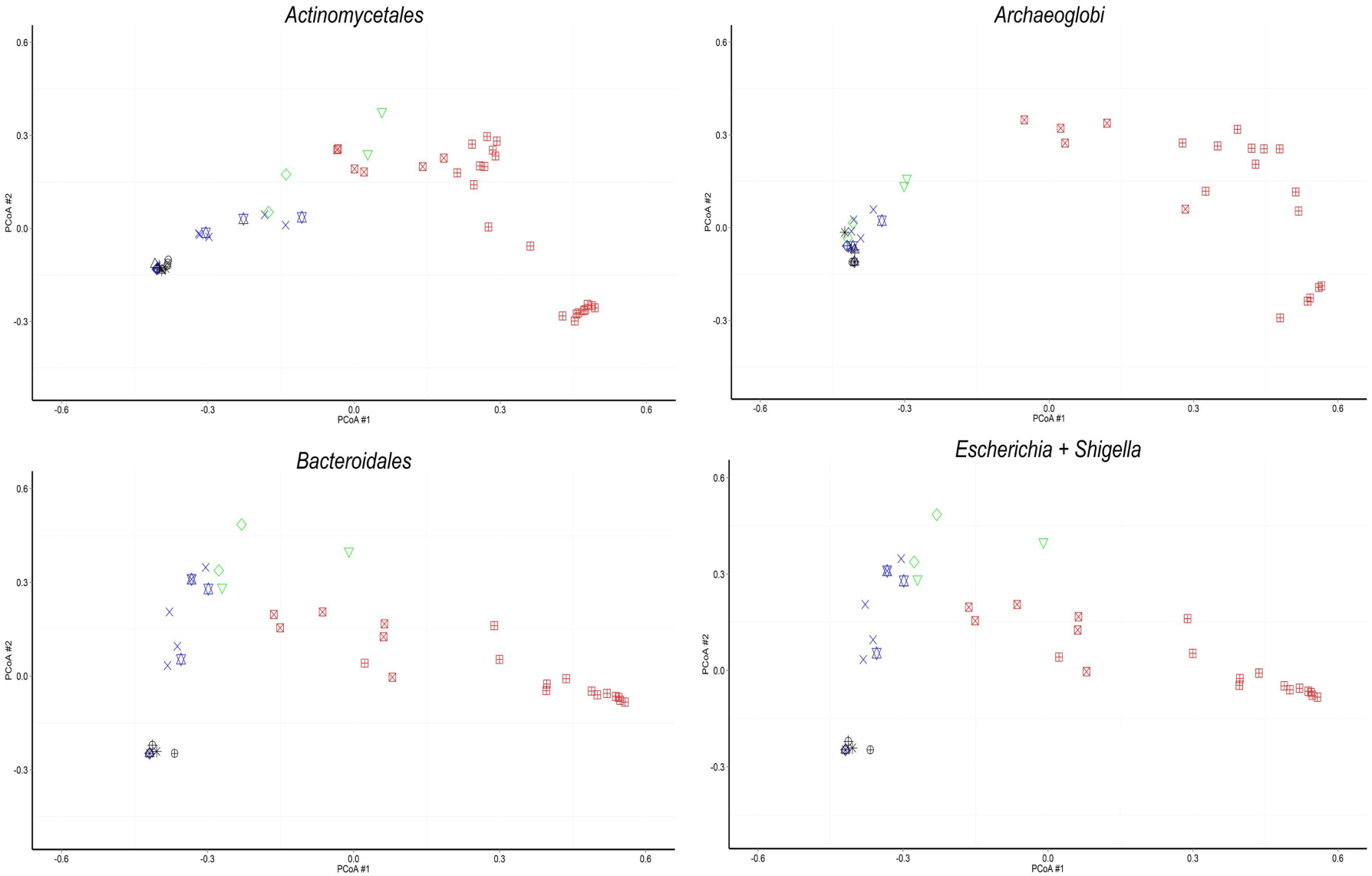
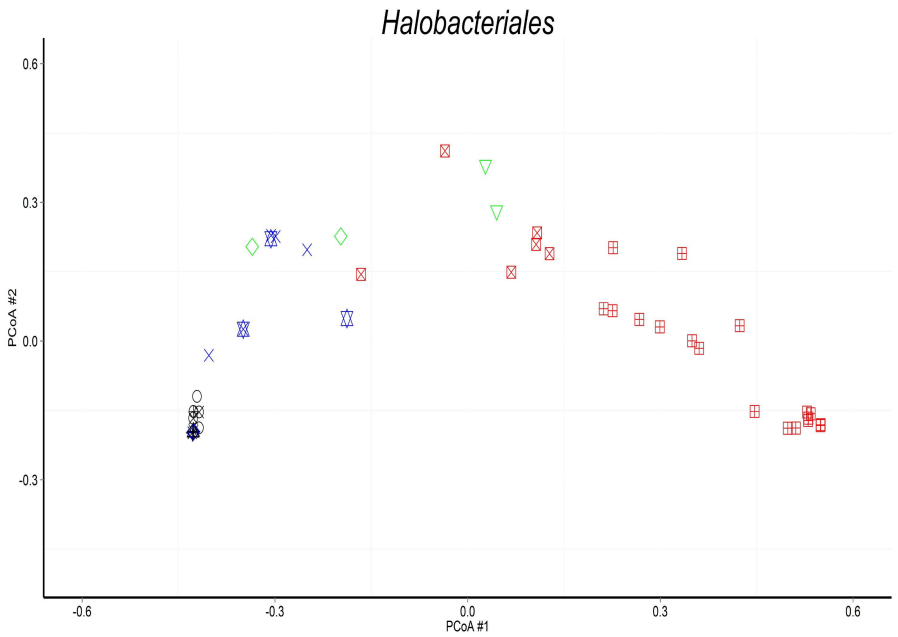
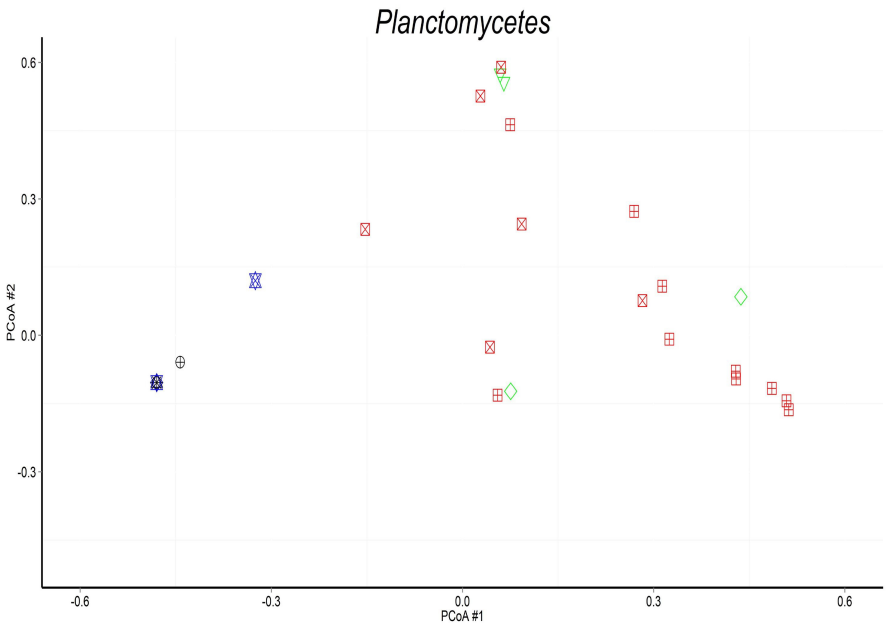
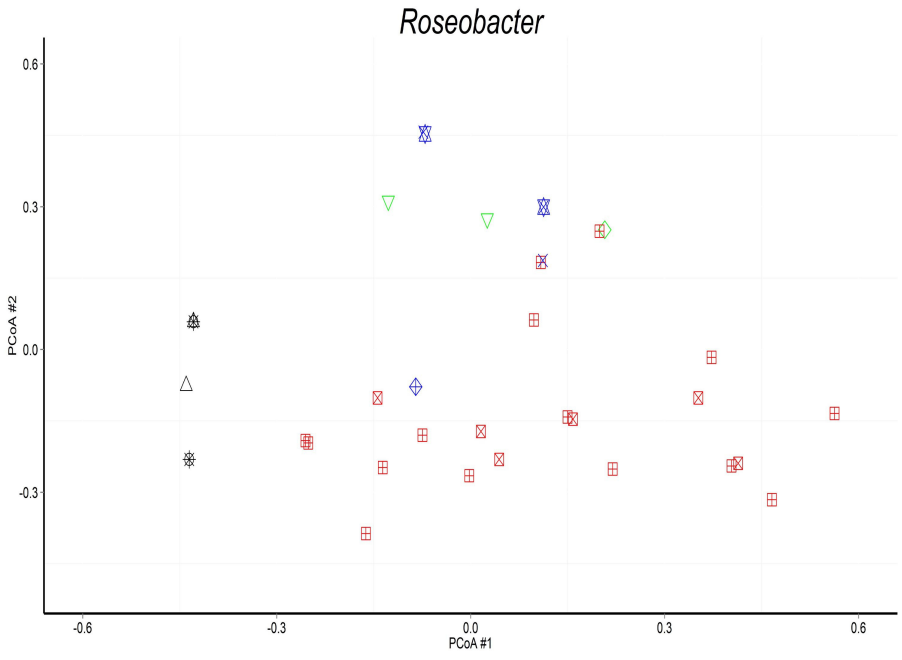


Figure 15. PCoA analysis using the RF distances between trees. (Actinomycetes, Archaeoglobi, Bacteroidales and Escherichia + Shigella). The plots have XY values of various colors and shapes. The details about the colors and shapes are shown in Figure 17.


Halobacteriales
PCoA #2
PCoA #1
0.6
0.3
0.0
-0.3
-0.6
-0.3
0.0
0.3
0.6


Planctomycetes
PCoA #2
PCoA #1
0.6
0.3
0.0
-0.3
-0.6
-0.3
0.0
0.3
0.6


Roseobacter
PCoA #2
PCoA #1
0.6
0.3
0.0
-0.3
-0.6
-0.3
0.0
0.3
0.6

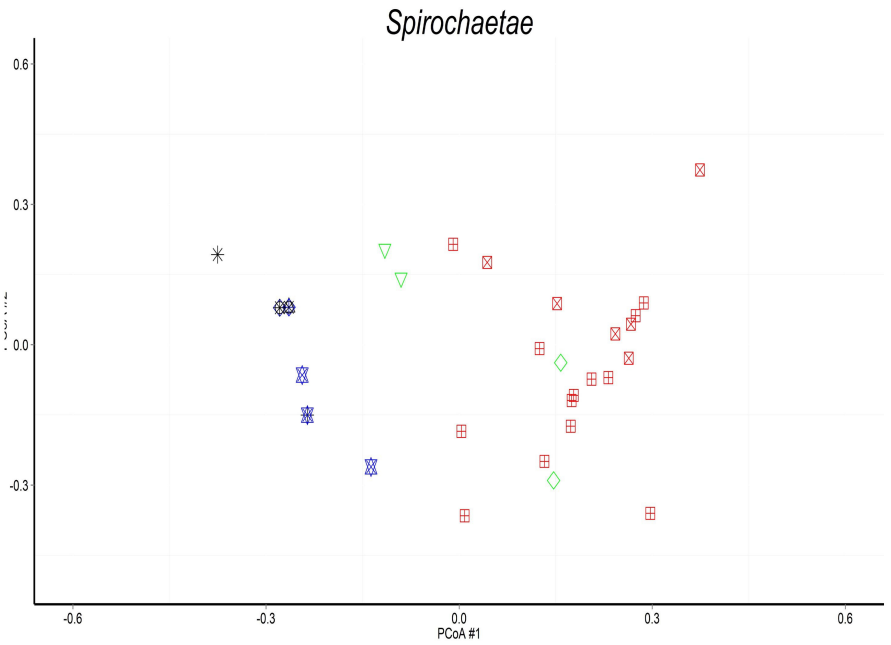

Spirochaetae
PCoA #1
0.6
0.3
0.0
-0.3
-0.6
-0.3
0.0
0.3
0.6

Figure 16. PCoA analysis using the RF distances between trees. (Halobacteria, Planctomycetes, Roseobacter, Spirochaetae). The plots have XY values of various colors and shapes. The details about the colors and shapes are shown in Figure 17.

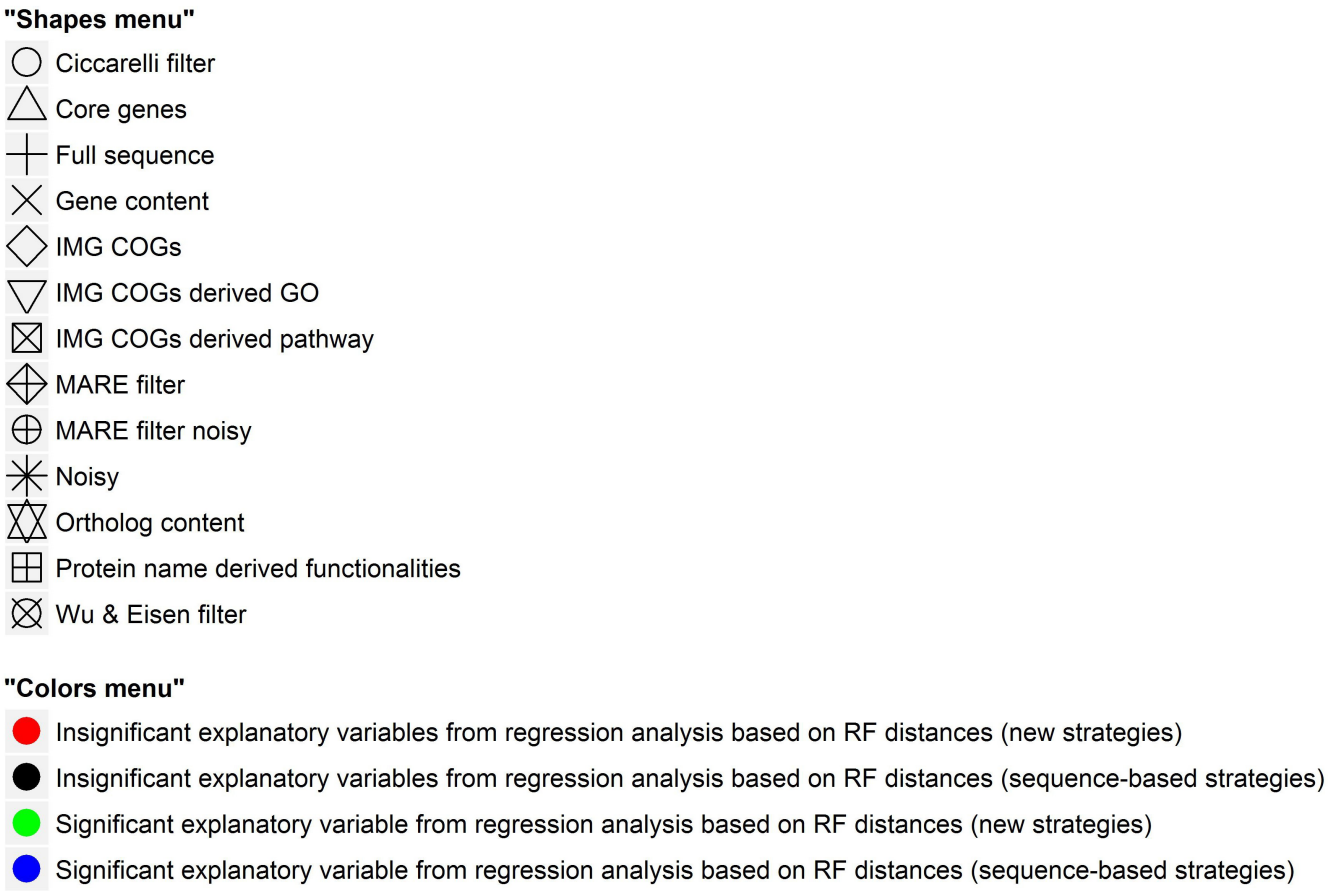


Figure 17. Colors and shapes menu for Figure 16 and 15. The colors indicate the insignificant and significant explanatory variables from regression analysis based on RF distances in the section 3.3.2. The shapes indicate the various sequence-based tree construction strategies and strategic variations using functionalities mapped from IMG COGs and protein names.

In Figure 15, the same information can be observed from the plots of different datasets. The reference trees from sequence-based supermatrix strategies (black colored) and functionalities based trees (red colored "double straight cut square" from Figure 17) mapped from protein names were distinctly grouped from each other on the PCoA plots with regard to RF distances. It indicates, both strategies are inferring topologically dissimilar phylogenies for the datasets shown. The sequence-based supermatrix strategies use potential number of concatenated alignment of informative nucleotide/amino acid

characters available for inferring phylogenies. In the case of functionalities mapped from protein names based strategies, the mapping system is not efficient to map correct functionalities from protein name. The weighting and applied filter schemes did not help to remove the ambiguity in mapping system completely. This must be the reason for topological dissimilarity between above both strategies.

In Figure 16, except *Halobacteriales,* there were no group formation observed in PCoA analysis plots for other remaining datasets. The remaining three datasets (*Spirichaetae, Planctomycetes, Roseobacter* clade) were small datasets containing less than 11 taxa. It indicates, trees with less taxa were not enough to visualize the difference between reference trees and reconstructed trees based on RF distance using PCoA analysis.

Among the trees reconstructed using 34 strategic variations, functionalities mapped from IMG COGs based trees show continuity in the topological similarities with sequence-based reference trees, especially GOs mapped from IMG COGs based trees and IMG COGs based trees in PCoA analysis of *Actinomycetales, Archaeoglobi, Bacteroidales, Escherichia + Shigella and Halobacteriales* of dataset. Moreover, IMG COGs and GOs mapped from IMG COGs based strategic variations (explanatory variable) showed low p-value in regression analysis with regard to RF distances between trees (dependent variable) in the section 3.3.2. In Figure 15 and 16, GOs mapped from IMG COGs and IMG COGs based trees were shown in green color.

Among the trees reconstructed using sequence-based strategies, gene-content and ortholog-content based strategies show continuity in the topological similarities with functionalities mapped from IMG COGs based trees. Moreover, these strategies (explanatory variables) showed low p-value in regression analysis with regard to (dependent variable) RF distances between trees in the section 3.3.2. In Figure 15 and 16, the above trees were shown in blue color.

From the above observations, functionalities mapped from protein names based strategies can not be used to infer the evolutionary relationship between taxa. The continuity of topological similarities between functionalities mapped from IMG COGs based strategies and gene/ortholog-content strategies were significantly increasing more when compare to supermatrix strategies. The gene/ortholog content and functionalities mapped from IMG COGs based strategies were following presence/absence approach. Even though MARE filtered sequence based strategy was observed with decreasing RF distances between trees in multiple regression analysis, the continuity of topological similarity with functionalities mapped from IMG COGs based strategies in the PCoA plot was not observed. Thus, the

presence/absence approach must be the key reason between the high continuity of topological similarity between gene/ortholog content and IMG COGs based functionalities based strategies.

3.4. DISCUSSION AND CONCLUSION

The genome association studies using genome encoded functionalities are important in genome evolutionary science [Huynen and Bork, 1998]. As such, 34 variations of three phylogenetic tree reconstruction strategies were formulated using shared functionalities. The screening was required to find a method which yields the most informative clues in phylogenetic inference. For this purpose, 34 variations and data sets used in each strategy were statistically tested for their suitability towards phylogenetic reconstruction and compared with sequence-based evolutionary trees in order to find the topological similarities between them.

Presence/absence COGs

The ambiguity raised during the mapping process from protein name to COGs were not fully solved by applying weights for the ambiguity. The presence/absence of COGs based strategy with 1.0 filter on assigned weights includes only the absolute presence of COGs in a genome. The above strategy shows increasing trend in treelikeness among eight dataset. This strategic variation is suitable to reconstruct phylogenetic trees when compared to other strategic variation applied filters on weights (0.25, 0.5, 0.75). All other presence/absence and abundance approaches apply reduced weights for the ambiguous COGs and include them in similarity calculation. But, the weighting schemes with applied filters failed to produce trees with high topological similarity with existing phylogenies. Thus, the weighted methods can not be used as an alternative strategy to reconstruct trees which show evolutionary information about organisms. It is concluded that phylogenetic reconstruction using the presence/absence strategy which considers the COG with absolute presence is the best approach when compared to other weighed approaches.

Functionalities mapped from protein names vs IMG COGs

From the regression studies and PCoA analysis with respect to the RF distances between trees in the section 3.3.3, it is concluded that strategies based on the functionalities mapped from protein names are not suitable to infer evolutionary relationship between taxa. Trees reconstructed using functionalities mapped from IMG COGs can be further considered with regard to topological similarities with other existing strategies.

Functionalities mapped from IMG COGs-based vs sequence-based trees

The gene/ortholog content based strategies use presence/absence methods instead of concatenated alignment to infer phylogenetic trees. The presence/absence based approach is a common approach between gene/ortholog sequence-based method and IMG COGs based strategic variations. The commonness must be the reason for the significant continuity in the topological similarity between above mentioned trees. Moreover, those strategies use homologous/orthologus entities from complete genome without any filters applied.

Advantages and application of functionalites based strategies

The advantage of using a shared functionalities based approach to infer phylogeny are: no need of computational resources to cluster orthologous groups of genes, COGs were obtained from already available resources, avoid uninformative data (e.g., hypothetical proteins and predicted proteins) for phylogenetic inference, the pathway-based trees infer the relationship between organisms in terms of their metabolic profile. Thus, it is useful in understanding the genomic evolutionary mechanisms of microorganisms in different habitats and their overall metabolic/biochemical variations [Hong et al., 2004].

Advantage and disadvantage of shared COGs based phylogenetic inference

The usage of shared COGs between genomes in phylogenetic reconstruction method is a straightforward and more simple approach than clustering orthologs in each genome using sequence similarity. Clustering orthologs is an intensive process in terms of computational memory usage and it is reduced in shared COG-based method. Because, COGs are precomputed clusters of orthologous group of proteins. The shared COGs based method trust the COGs annotated by IMG for each organisms as true orthologous clusters. Thus, this method depends on the acquired data.

Advantage of shared pathways based phylogenetic inference and future possiblity

The total gene number, non-essential, non-conserved sequences and size of genomic DNA in an organism affect the sequence-based phylogeny reconstruction methods. The metabolic pathway is a representative of functionally linked genes which are necessary for the survival of an organism. In this connection, this strategy minimizes the effect of the non-conserved and uninformative details in a

genome for phylogenetic inference. The pathway-based trees represent both the changes in genetic contents and metabolism [Hong et al., 2004]. The evolution of entire pathways can be compared to the evolution of single genes (enzymes) that are related to the pathways.

Advantage of shared GOs based phylogenetic inference and future possiblity

The GO based tree reconstruction approach considers the functional properties and characteristics of genes. The usage of functional information from a gene instead of a gene sequence is a straightforward approach and reduces computational resource as well as time. This approach also has the features of above shared functionalities based methods such as the effect of uninformative details in a gene/genome for phylogenetic inference can be reduced and the evolution of entire gene ontologies can be compared to the evolution of single molecular characteristics/biological process.

4. EVOLUTIONARY CORRELATION BETWEEN FUNCTIONALLY LINKED CHARACTERS

4.1. INTRODUCTION

A character can be described as a distinguishable features or functions possessed by organisms. Some characters quickly evolve than others and some in dependent manner with others. By studying the nature of these evolutionary processes for a character, the relationship between correlated evolution and functional role of characters can be determined. In this chapter, an approach had been conducted to infer the correlated evolution of functionally linked characters (genes/enzymes) across a group of organisms. The combined approach uses a phylogenetic tree of a group of organisms and presence/absence of characters e.g. genes/enzymes across a set of organisms. With a phylogenetic tree and a distribution of character states, the evolution of characters can be traced over a phylogenetic tree using the BayesTrait and the correlated evolution of two or more characters can be inferred.

For example, *Spirochaetes* rely on increasing their number of motility genes in for increasing their number of flagella. The correlation between the proportion of motility genes and the average number of flagella reported for *Spirochaetes* was significantly high [Abt et al., 2013]. It infers that the number of motility genes were gained in organisms for the common functional reason and they are directly proportional to the number of flagella. In this case, the chance of the correlated evolution of functionally linked motility genes is high. The evolutionary study of functionally linked genes involved in a specific pathway (e.g. motility pathways) will provide insight into the environmental adaptation and gain of new function in organisms.

4.1.1. CHARACTER EVOLUTION

In each branch of a phylogenetic tree, a character initially present may disappear or a character initially absent may gain. The changes along the branches are called character state transition [Harvey and Pagel, 1991]. A particular pattern of observed transitions between states of characters throughout the phylogeny provides evolutionary changes of characters across organisms. In this manner, a phylogeny and set of characters with their states in a group of organisms are used for studying correlated evolutionary study of characters. The evolutionary correlation of genes, enzymes, pathways, and genomic features were studied in this chapter. The correlation studies are classified based on the type of characters: discrete characters and continuous characters.

1. Discrete characters

A character which varies in finite states is called a discrete character. The different shapes of bacterial cells are discrete characters (e.g. round, oval). Ridley's method and Pagel's method are example methods to study evolution correlation between two or more discrete characters.

Ridley's method

Ridley's method considers a given phylogeny as a true phylogeny. It maps character states on a tree using the parsimony concept. Once the character states are mapped over higher nodes in a tree, the number of state transitions are tallied. The character states can be tallied as two-by-two contingencies. The hypothesis of independent or dependent correlation between characters can be tested using the chi-square statistics. A significant p-value from the chi-square test is an evidence that transitions are dependent each other [Harvey and Pagel, 1991].

Pagel's method

In 1994, Pagel introduced the first likelihood-based correlation method for discrete characters. If two characters are evolving in a correlated manner, the state of each character will affect the probability of a change of the other one. There are four possible state combinations for two characters: 00, 11, 01, and 10. Pagel used α and β rates for forward and backward changes for the first character states, and γ and δ rates for changes to the second character states [Pagel, 1994].

The combinations of the rates of change with regard to the character states in Pagel's discrete character comparative method are shown in Table 4.

Table 4. Character state transitions – Pagel's method

To: / From:	00	01	10	11
00	--	γ_0	α_0	0
01	δ_0	--	0	α_1
10	β_0	0	--	γ_1
11	0	β_1	δ_1	--

The likelihood can be calculated from the transition probabilities of character state changes. The

likelihood ratio can be calculated by comparing the likelihood of the four restricted parameter values (if the positions of α, β, γ, δ do not affect their values, the positions α_0 and α_1 can be reduced. The number of the parameter will be 4 instead of 8) to the eight unrestricted parameter values. With the applied restrictions on the parameters, the likelihood of dependent and independent character state changes between two characters can be calculated [Pagel, 1994].

Pagel's discrete character correlation analysis method was used in this study for discrete characters instead of Ridley's method. Ridley's method just provides the information that character transitions are dependent each other. Using the Pagel's method, a likelihood ratio (LR) test can be performed with respect to the likelihood of dependent vs independent character state changes between two characters. Using the LR test, the probability of correlated evolution between two characters can be calculated.

2. Continuous characters

Continuous characters are characters with infinite number of character states between two given points. In the case of bacterial cell's sizes were measured, the size would be continuous as there are an unlimited number of possibilities even though it is measured between 0.2 and 3 microns.

Random walk model

The random walk is a mathematical phenomenon which explains a path that consists of a succession of random steps [Pearson, 1905]. The random walk model in evolutionary science assumes that a continuous character evolves across a group of organisms in a phylogeny in random fashion. It applies further assumptions and conditions. The covariance of a character state at the tips of phylogeny can be calculated through the variances contributed by the branches that are shared by the path from the root up to the most recent common ancestor of the tips. From this covariance, the likelihood can be calculated if the character evolves according to the assumptions made. This covariance can be used to compare the evolution of two characters across a phylogenetic tree [Felsenstein, 2004].

In this study, the random walk model was applied to find the correlation between continuous characters in evolution.

4.1.2. CHARACTER CORRELATION

BayesTraits is a software package for finding evolutionary correlations between two characters across a group of organisms where a phylogeny or sample of phylogenetic trees are available. It can be applied

to discrete characters or continuous characters [Pagel et al., 2004]. BayesTraits has two models for studying evolution of discrete characters and three models for continuous characters.

Discrete characters

The BayesTraits models for studying evolution of discrete character use one or more phylogenetic trees and a pair of characters with the (presence or absence) character states in a group of organisms in order to find evolutionary correlation between two characters. The first model assumes that the characters were evolved in dependent manner and second one as independent manner.

In the first model, both characters are dependent on each other. It has eight rate parameters (α_0, α_1, β_0, β_1, γ_0, γ_1, δ_0, δ_1). The parameters and their transitions are mentioned in the Table 5.

The independent model has four rate parameters (α, β, γ, δ) with the assumption that the two characters were independently evolved. The parameters and character transitions are mentioned in Table 5.

Table 5. Character state transitions – dependent and independent models

	Dependent model					Independent model			
To: / From:	00	01	10	11		00	01	10	11
00	--	γ_0	α_0	0		--	γ	α	0
01	δ_0	--	0	α_1		δ	--	0	α
10	β_0	0	--	γ_1		β	0	--	γ
11	0	β_1	δ_1	--		0	β	δ	–

Continuous characters

The standard constant-variance random walk model was used in this study for finding correlation between continuous characters. The BayesTraits continuous random walk model uses a phylogenetic tree and a pair of characters with the continuous character state data across a group of organisms [Pagel et al., 2004]. Using the above model, a likelihood ratio (LR) test can be performed with respect to the likelihood obtained from the same model with an assumption of true character correlations and false character correlations between two continuous characters. Using the LR test, a probability of correlated evolution between two continuous characters can be calculated.

4.2. MATERIALS

4.2.1. DATASETS

In each dataset, the bacterial/archaeal genomes were grouped based on same taxonomic rank. The datasets used in this study were *Planctomycetes*, *Archaeoglobi* + outgroups, *Escherichia* + *Shigella, Halobacteriales* + outgroups, *Bacteroidales, Roseobacter* clade, *Spirochaetae,* and *Actinomycetales*. Information about the organisms in each dataset is provided in Table S1 from the supplementary materials section. These datasets were used to optimize the threshold value to get best one for finding significantly correlated genes.

The *Spirochaetae* dataset with 29 organisms was used to find evolutionary correlation between functionally linked genes. The *Rhodobacteraceae* dataset with 87 organisms was used to find evolutionary correlation between functionally linked enzymes, pathways and genomic features. Organisms in those datasets are listed in tables S2 and S3 in the supplementary materials section.

Tree file

The nexus format of the rooted tree file is required to execute BayesTraits. The trees were rooted using the software package Phylip retree [Felsenstein, 2005]. The midpoint rooting method was used for rooting trees. The trees used in this study were whole-genome sequence-based trees. The trees were created by my colleague Carmen Scheuner using the methods explained in Figure 14.

Characters file (EPF file format)

The main EPF character file contains n columns. The first row contains the number of characters, the first column contains the organism names or indices, the second column contains character states (presence/absence) of first character across all organisms and the n^{th} column contains character states (presence/absence) of $(n-1)^{th}$ character across all organisms. BayesTraits software requires three column file, where first column is organism names and other two columns contain characters states across organisms. The EPF format is convenient to automatically generate three column file for BayesTraits execution between two characters. Thus main character file is used as the EPF format.

4.2.2. TYPES OF CHARACTERS

Different types of biological characters were used in this study and sourced from various resources.

Discrete characters – genes

Annotated genes and enzymes from microbial genomes were used as characters in discrete BayesTraits correlation analysis. The presence/absence details of genes in a group of organism were considered as character states. The presence/absence details of gene content in a group of organisms were obtained from my colleague Carmen Scheuner. She obtained the data using the tribeMCL clustering approach mentioned in Figure 14 for the *Spirochaetae* dataset with 29 organisms.

Discrete characters – enzymes, pathways

The presence/absence details of enzymes and pathways for *Rhodobacteraceae* datasets were sourced from the collaborating research team of Prof. Dietmar Schomburg, TU-Braunschweig, Germany. The presence/absence of enzymes were identified using EnzymeDetector by his team. The EnzymeDetector is a tool used to obtain the organism-specific enzyme annotations from various resources. This tool compares and evaluates the previously assigned enzyme functions from the annotation databases. It supplements the annotations with its own function prediction using the sequence similarity analysis over manually curated organism-specific enzyme information from BRENDA [Schomburg et al., 2013], [Quester and Schomburg, 2011]. The mapped pathways (Only for *Rhodobacteraceae* dataset) with more than 75% of its enzymes availability in an organism were considered as pathway presence and others as absence.

Continuous characters – genomic features

The genomic features (continuous characters) of *Rhodobacteraceae* dataset with 87 organisms were collected from the IMG server by my colleague Oliver Frank. The genomic features used in this correlation study are number of proteins in a genome, length of a genome, GC content in a genome, proportion of genes grouped in the COG groups such as “information processing and storage”, “cellular processing and signaling”, “metabolism”, “genes categorized as poorly characterized”, proportion of genes grouped in the COG categories in a genome, number of SSU in a genome, number of LSU in a genome, number of CRISPRs (Clustered Regularly Inter spaced Short Palindromic Repeats), number of tRNAs, number of ribosomal proteins, number of transporter proteins and number of ABC (ATP binding cassettes). The abbreviations are mentioned in the Table S5.

4.3. METHODS

4.3.1. BAYESTRAITS EXECUTION

All characters from the character file were paired with other characters to find evolutionary correlation between each pair by the implemented pipeline using the Ruby script. As BayesTraits can only accept a pair of characters with presence/absence details as direct input, a pipeline had to be implemented to automatically execute BayesTraits for each characters pair from the character file. The model selection for different character data (discrete and continuous characters) is explained in the sections below.

4.3.2. LIKELIHOOD RATIO STATISTICS

The likelihood ratio (LR) was calculated using the formula LR= 2(log-likelihood(model with best parameters) – log-likelihood (model with worst parameters)). For discrete character pairs, the model with best parameters is the BayesTraits model which assumes that the characters evolved in dependent manner and the model with worst parameters is the BayesTraits model which assumes that the character evolved in independent manner. For continuous character pairs, the model with best paramenters is the BayesTraits random walk model with the assumption of true correlation and the model with worst parameters is the BayesTraits random walk model with the assumption of false correlation.

The likelihood ratio is nominally distributed as χ^2 with the degrees of freedom 1 using the chi-squared distribution probability function. The p-values were calculated from the LR test for each character pair. Various threshold values were applied (0.1, 0.075, 0.05, 0.025, 0.01, 0.005 and 0.001) to the p-values obtained from the chi-squared distribution for the eight datasets mentioned in Table S1. The threshold value was optimized (described in the section 4.4.1) for discrete characters and applied to the *Rhodobacteraceae* dataset with 87 organisms and the *Spirochaetae* dataset with 29 organisms. The likelihood ratio statistic yields the three column file with first two columns containing characters and third column containing the p-values from the chi-squared distribution analysis for each character pair of characters from the same row.

4.3.3. CLUSTERING CHARACTERS

The characters were clustered using the MCL algorithm [van Dongen, 2000]. The clustering process needs three column file, where first and second column with characters and third column with

similarity value between the characters in same row. The significant p-values ($p < 0.05$) obtained from the chi-squared distribution probability function for character pairs (as mentioned in the above section) were converted to negative logarithmic values and used to specify the weight of the arc connecting the two nodes. The negative logarithms calculated from the p-values were used as a similarity value between characters. The clusters were formed with respect to the predefined inflation value of 2.0 which determines the granularity of clusters. The characters which did not significantly pair with any other character in the chi-squared distribution probability function mentioned in the above section are singlets. Those singlets were used in the clustering process by including the same character in first two columns with high negative logarithm value as a similarity between them.

4.3.4. EVOLUTIONARY CORRELATION BETWEEN FUNCTIONALLY LINKED GENES

The strategy of grouping genes which evolved in correlated fashion and functionally linked has four stages: estimation of gene pairs with significant evolutionary correlation, mapping genes to their functional groups (pathways and COG categories/groups), clustering evolutionary correlated genes, grouping genes which are functionally linked and correlated in evolution. The *Spirochaetae* dataset with 29 organisms was used in this method.

The models for evolutionary study of discrete data from BayesTraits were selected (as mentioned in the section 4.3.1) and likelihood value from those models were used to calculate the LR between each gene pairs. The chi-squared distribution function provides the p-value for LR ratio obtained for each gene pairs in the dataset (as mentioned in the section 4.3.2). As genes are considered as discrete characters, the work flow described in Figure 19 represents the BayesTraits execution for genes too.

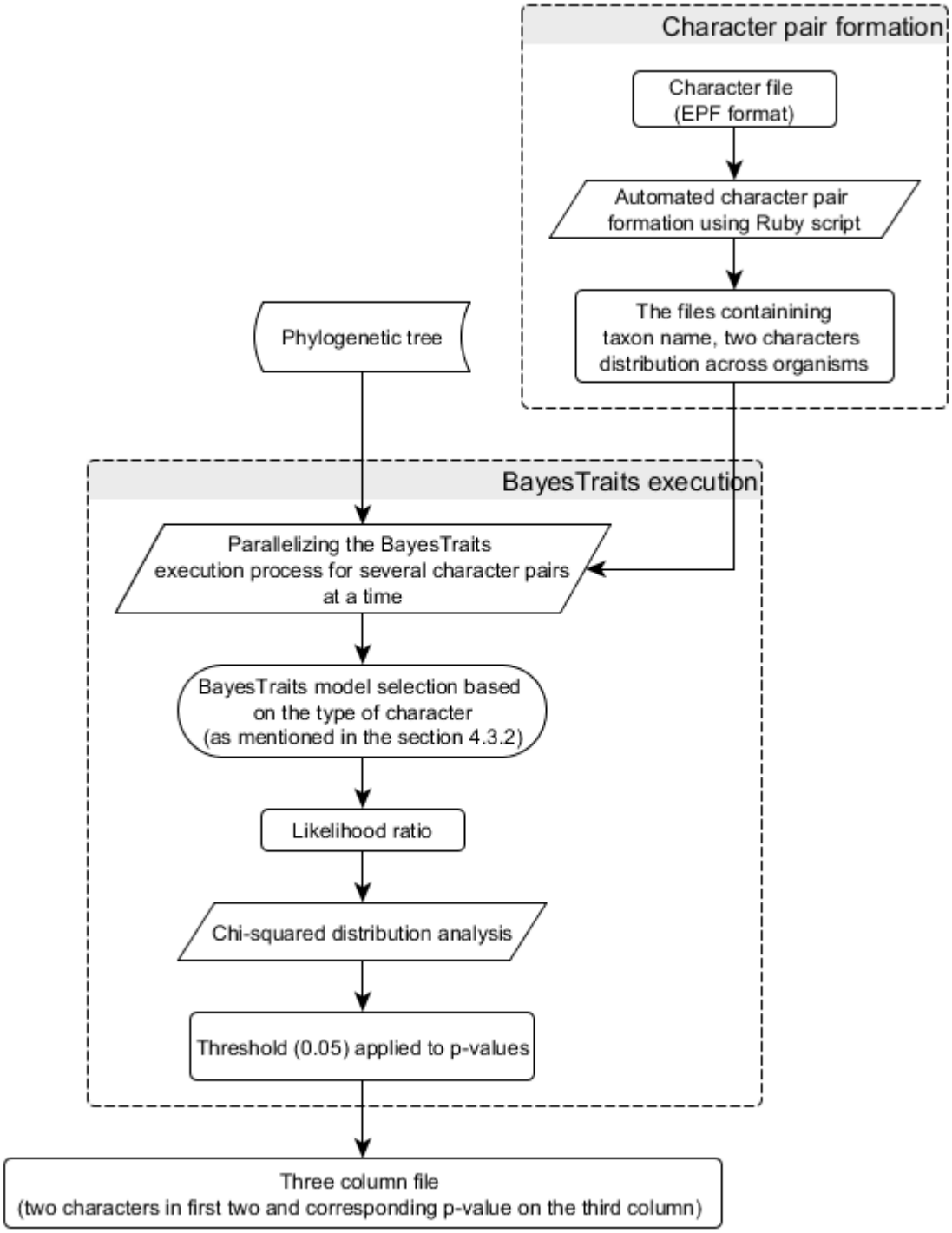


Figure 18. The flow chart describes the parallelised BayesTraits execution, model selection for discrete characters and p-value estimation from the LR obtained from selected models. The parallelogram indicates the methods, elongated oval shape indicates the selection of model.

Mapping system – COG categories/groups

The COGs were mapped from genes using "gene – COG" cross-references obtained from the IMG server. The COG categories/groups were mapped from COGs using the locally implemented PostgreSQL database contains relations for COG categories/groups. In Figure 19, COG categories/groups mapping system is described.

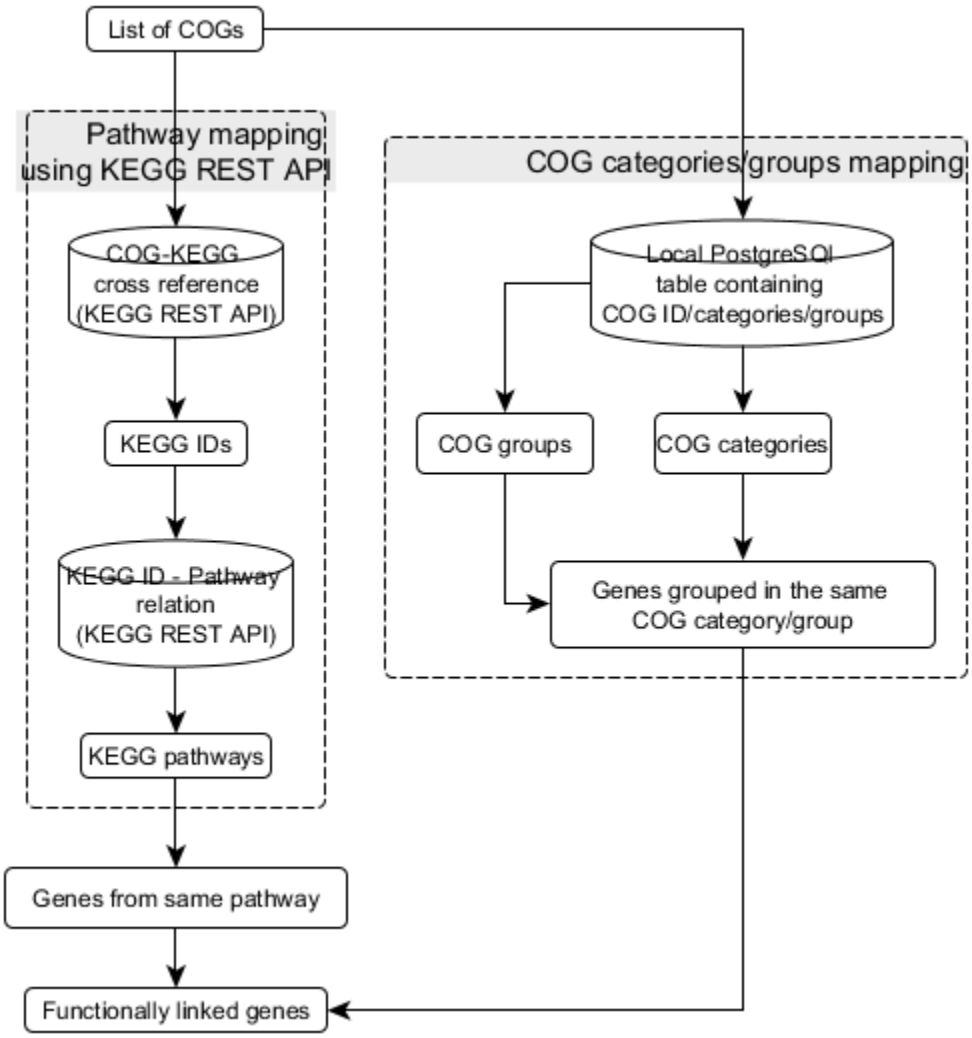


Figure 19. A map to COG categories/group mapping and pathway from COGs. The cylinder indicates the local/remote database.

Mapping system – pathways

KEGG pathways were mapped from COGs using LinkDB. LinkDB is a platform used to retrieve information from the KEGG server and links to other databases. The cross-referential information of a given entry in a given database can be obtained by invoking the search command against LinkDB. It contains original links provided by each database and indirect links (e.g. accession between Genbank and EMBL). The links between databases were pre-computed in LinkDB [Fujibuchi et al., 1998].

The relationships between remote databases were established using LinkDB through the Representational State Transfer (REST). REST is an architecture style developed based on the already existing design of HTTP [Fielding and Taylor, 2002]. REST is a method to access the remote resource entities which are referenced with a global identifier (e.g. a Unified Resource Identifier). In order to access these resources (online databases), components of the network (client and server) communicate through a standardized interface (e.g. HTTP) and exchange representations of resources. It uses the normal get/request operations to obtain the cross-references. LinkDB uses the DBGET network distributed database system [Fujibuchi et al., 1998]. To access the KEGG server, the REST operations have a URL structure which contains operations and arguments at the end (e.g. http://rest.kegg.jp/<operation>/<argument>/<argument2>, where “list”, “find”, “get”, “conv” are some examples operations, database or data entries which can be used as arguments). For example, the URL for obtaining cross-references from NCBI-GI to KEGG ID is “http://rest.kegg.jp/conv/genes/ncbi-gi:3113320”. LinkDB was used in this study to map pathways. It is the sophisticated way, if the internet connection speed is good enough. An automated REST mapping system was developed using the Ruby script. As SQL updates from KEGG resource had been shut down, KEGG REST API is an alternative method to access KEGG resource. The pathway mapping system is described in Figure 19. The complete work flow of finding evolutionary correlation between functionally linked genes were shown in Figure 20.

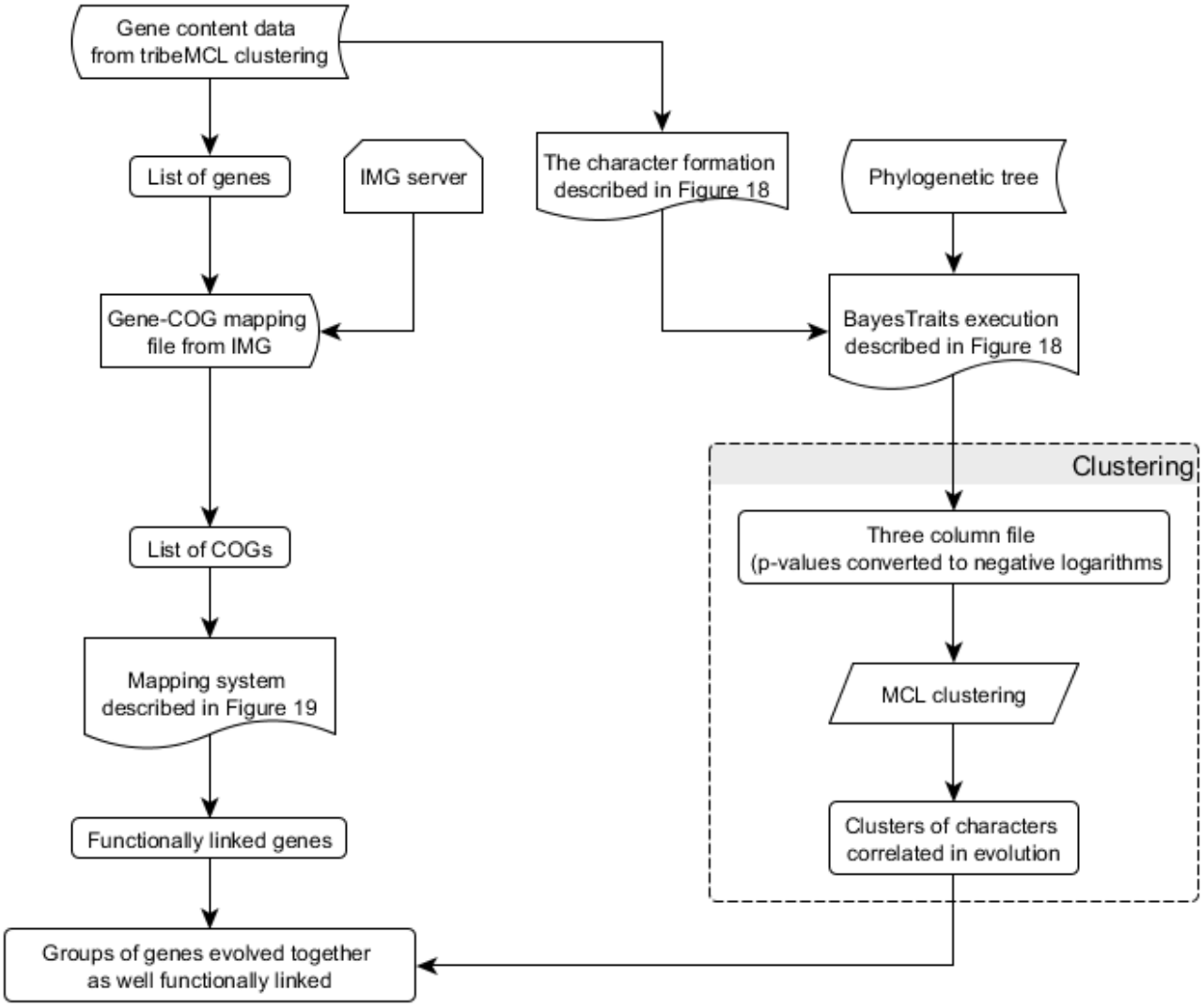


Figure 20. The evolutionary correlation study of functionally linked genes. The BayesTraits software execution, clustering, mapping of pathways and COG categories/groups and grouping functionally linked genes are collectively shown along with Figure 18 and 19.

4.3.5. EVOLUTIONARY CORRELATION BETWEEN FUNCTIONALLY LINKED ENZYMES

The strategy of grouping enzymes which evolved in correlated fashion and functionally linked has four stages as mentioned in the section 4.3.4. The *Rhodobacteraceae* dataset with 87 organisms was applied in this method. The complete work flow of finding evolutionary correlation between functionally linked enzymes are shown in Figure 21.

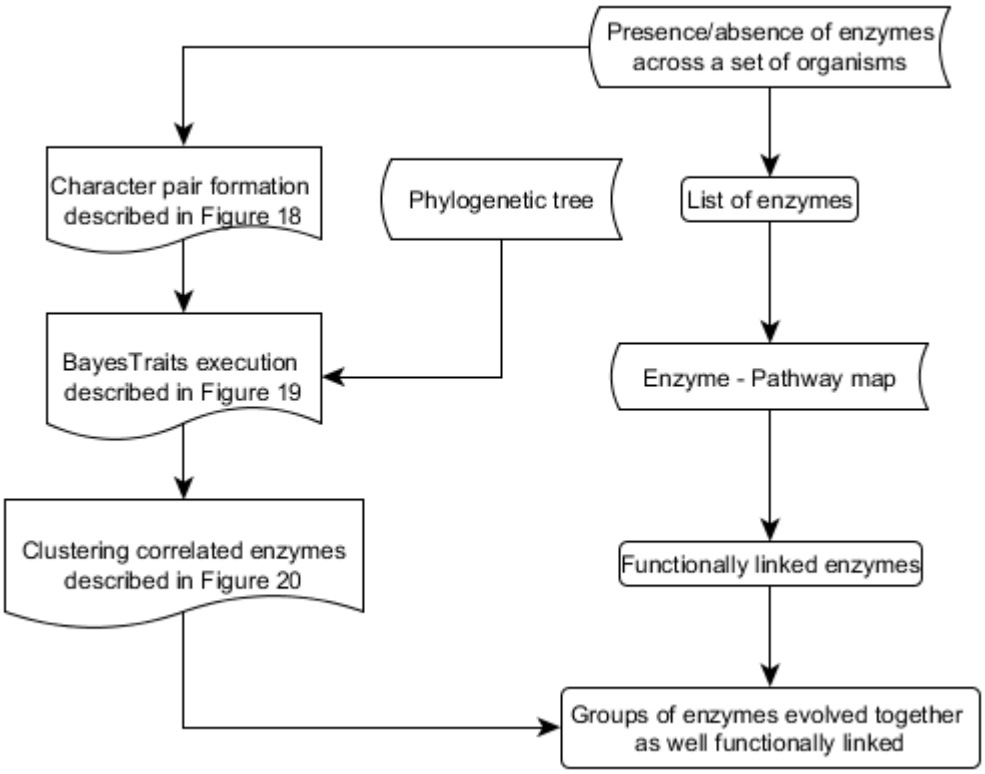


Figure 21. The evolutionary correlation study of functionally linked enzymes. The BayesTraits software execution, clustering, mapping of pathways and grouping functionally linked genes are shown. As enzymes were considered as discrete characters, the BayesTraits execution with the models indicated in Figure 19 were followed.

The distribution of correlated enzymes which involved in the same pathways were reported across 87 organisms of *Rhodobacteraceae* dataset in the section 4.4.4.

4.3.6. EVOLUTIONARY CORRELATION BETWEEN PATHWAYS

The strategy of finding correlation between pathways has two stages: the estimation of pathway pairs with significant evolutionary correlation and clustering evolutionary correlated pathways. The *Rhodobacteraceae* dataset with 87 organisms were applied in this method. The complete work flow of finding evolutionary correlation between pathways are shown in Figure 22.

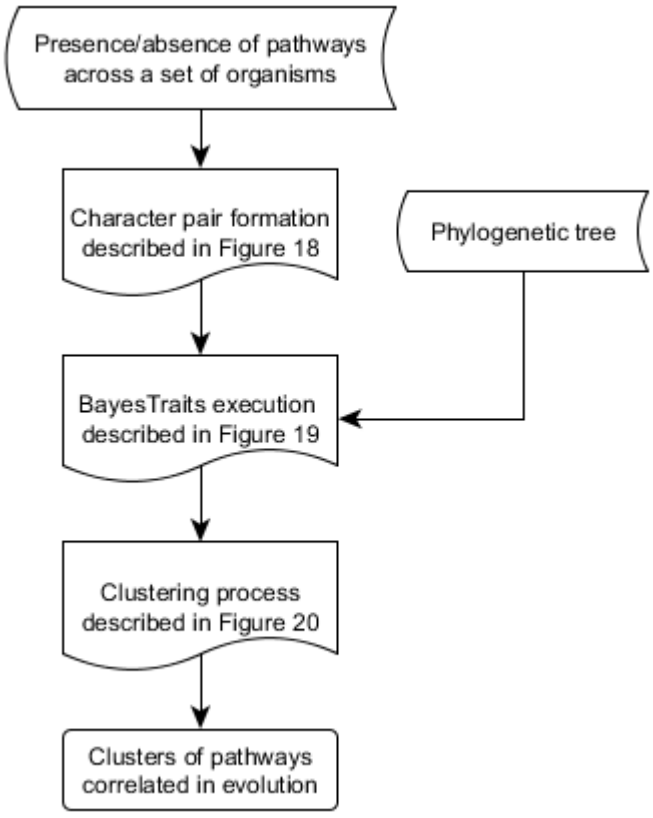


Figure 22. The evolutionary correlation study of pathways. The BayesTraits software execution and clustering of pathways are shown. As pathways were considered as discrete characters, the BayesTraits execution with the models indicated in Figure 19 were followed.

4.3.7. EVOLUTIONARY CORRELATION BETWEEN GENOMIC FEATURES

The strategy of clustering genomic features has two stages: the estimation of genomic feature pairs with significant evolutionary correlation and clustering evolutionary correlated genomic features. The *Rhodobacteraceae* dataset with 87 organisms were applied in this method. The complete work flow of finding evolutionary correlation between genomic features are shown in Figure 23.

The BayesTraits continuous evolutionary model was used to find correlation between genomic features. The threshold value 0.01 was applied to p-values obtained from the chi-squared distribution analysis using LR values between genomic features.

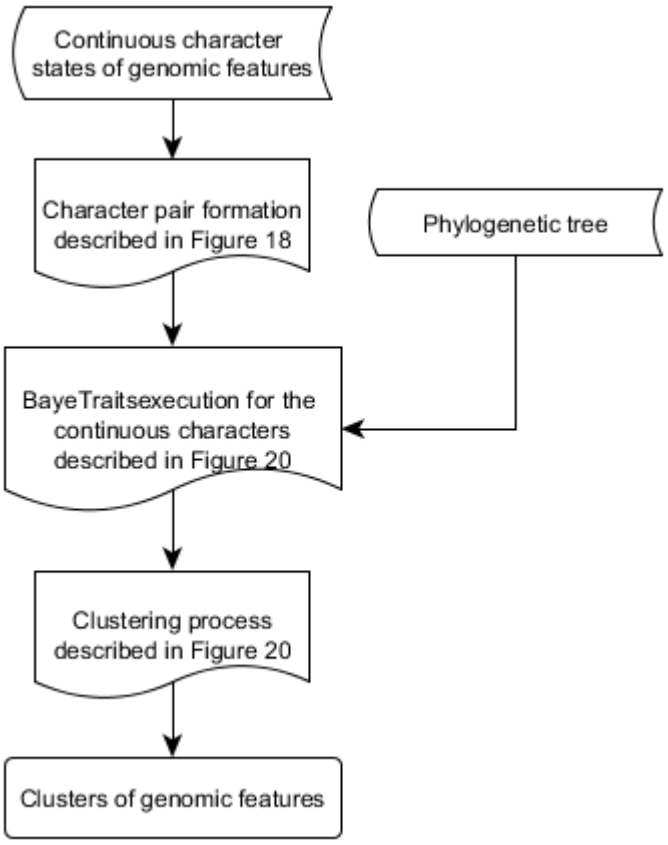


Figure 23. The evolutionary correlation study of genomic features. The BayesTraits software execution and clustering of genomic features are shown.

4.3.8. PEARSON'S CHI-SQUARED TEST

The Pearson's chi-square test of independence is normally used to test the independence or dependence of two variables. Two categories of variables can be tested using Pearson's chi-squared test [Pearson, 1990].

If variables of two categories were tabulated as row and column with observed values as contingencies, the expected frequency of the model can be calculated using the formula: Expected frequency = $\frac{Row\,total \times Column\,total}{Grand\,total}$. The dependence/independence of both categories (goodness of fit) for observed values can be estimated with the formula: $\frac{Observed\,value - Expected\,frequency}{Expected\,frequency}$. The positive goodness of fit indicates that the observed value is fitted well for the model and vice versa.

The Pearson's chi-squared test was applied in the results analysis like threshold optimization for LR statistics, grouping evolutionary correlated genes which are also classified under same COG groups/categories, grouping evolutionary correlated genes which involved in the same pathway and grouping evolutionary correlated enzymes which involved in the same pathway.

4.3.9. THRESHOLD OPTIMIZATION

The gene pairs with significant p-value of LR statistics obtained from the likelihood of BayesTraits dependent/independent models were hypothesized as correlated gene pairs. Using eight datasets (listed in Table S1), the categories applied in Pearson's chi-squared test for various threshold values (0.1, 0.075, 0.05, 0.025, 0.01, 0.005 and 0.001) are: gene pairs observed from same pathways and significantly/insignificantly correlated gene pairs with respect to applied threshold values.

Table 6 shows an example contingency table for two categories mentioned above for *Roseobacter* clade dataset for the threshold value $p < 0.05$. The numbers in bracket represent the field index of four fields in contingency table.

Table 6. Contingency table with two categories for Roseobacter clade dataset

	Gene pairs observed in same pathway	**Gene pairs not observed in same pathway**
Gene pairs with p < 0.05 from LR statistics (significance)	30967 (1)	306443 (3)
Gene pairs with p >= 0.05 from LR statistics (insignificance)	234086 (2)	2862015 (4)

As seen in Table 7, gene pairs involved together in the same pathway as well as evolved in correlated fashion (field index 1) and gene pairs not involved in same pathway as well as not evolved in correlated fashion (field index 4) provide the list of gene pairs related to each other by means of both correlated evolution and pathway involvement. The other two fields provide list of gene pairs which are not related each other by means of correlated evolution or pathway involvement. Thus, the goodness of fit obtained for the values in the field 1 and 4 provide the dependence/independence between pathway involvement and correlated evolution of gene pairs.

The Pearson's chi-square test produced expected frequencies for four fields of contingencies (mentioned as an example in Table 6) created for eight datasets with seven different threshold values. The goodness of fit for fields 1 and 4 were determined using the formula

$$\frac{(Observed(1)+Observed(4))-(Expected(1)+Expected(4))}{Expected(1)+Expected(4)}$$.

4.3.10. CHI-SQUARED TEST: GENE CLUSTERS VS COG GROUPS/CATEGORIES AND PATHWAYS

A chi-squared test was conducted for the dataset listed in Table S2 (*Spirochaetae* dataset with 29 genomes) with regard to categories correlated gene clusters versus genes classified under COG groups/categories and pathways. The COG groups/categories and pathways were mapped using the locally implemented pipeline (described in Figure 19). Contingency tables were created with the correlated gene clusters in columns and rows as COG groups/categories or pathways. For example, a contingency table for COG groups (*Spirochaetae* dataset with 29 genomes) *vs*. correlated gene clusters is shown in Table 9 for the number of genes in each contingency.

Table 7. The contingency table of the COG groups vs. correlated gene clusters for the Spirochaetae dataset with 29 genomes

COG groups	**Number of genes in cluster 1**	**Number of genes in cluster 2**	**...**	**Number of genes in cluster 330**	**Singlets**
Poorly characterized	82	110	...	0	488
Cellular processes and signaling	62	95	...	0	280
Information storage and processing	58	37	...	1	338
metabolism	128	75	...	1	436

The goodness of fit values were generated for the number of genes in each contingency. If the specific cell in the table shows the positive goodness of fit value, it indicates particularly these clusters of genes grouped in specific functional group (COG group/category or pathway) were correlated in evolution together. The purpose of this chi-squared test is finding specially interesting clusters of evolutionarily correlated genes which are functionally linked either in same COG group/category or same pathway. For example, a cluster of evolutionarily correlated genes involving in motility pathways were found in *Spirochaetae* dataset.

4.3.11. CHI-SQUARED TEST: ENZYME CLUSTERS VS PATHWAYS

In the correlation study of functionally linked enzymes, the *Rhodobacteraceae* dataset with 87 genomes and 1579 enzymes were used. 40 evolutionarily correlated enzymes were clustered.

Contingency table was created with the enzyme pairs correlated in evolution versus involvement of enzyme pairs in pathways and shown in table 8.

Table 8. The contingency table of the correlated enzyme pairs vs enzymes involved in same pathway. The numbers in bracket indicate the field index number for four fields contingencies.

	Enzyme pairs observed in same pathway	**Enzyme pairs not observed in same pathway**
Correlated in evolution	389 (1)	262541 (3)
Uncorrelated in evolution	451 (2)	833296 (4)

The goodness of fit value was calculated using the Pearson's chi-squared tests for the four fields in Table 8. The goodness of fit for field 1 and 4 showed positive values of 0.9315 and 0.0002. It indicates that observed values in field 1 and 4 are higher than expected frequency. It infers the dependency between evolutionary correlation and functional involvement of enzyme pairs in the *Rhodobacteraceae* dataset.

4.3.12. CHARACTER STATE RECONSTRUCTION

The evolutionary correlations of several genomic characters with one character (e.g., pathways vs living environments of *Rhodobacteraceae* like marine/non-marine) were further studied. 41295 genes, 57802 orthologs, 1728 enzymes, and 323 pathways (as described in section 4.2.2) of a *Rhodobacteraceae* dataset with 108 genomes were studied to identify evolutionary correlations with the living environments. The living environments of *Rhodobacteraceae* considered in this study were marine and non-marine. The discrete character correlations were identified using the BayesTraits software and Pagel's method was chosen. The required phylogeny was created by my colleague Carmen Scheuner using the concatenated core-genes alignment method (described in section 3.2.7). The character matrices with presence or absence details of characters in 108 organisms were used to find evolutionary correlations with regard to the living environments of *Rhodobacteraceae*. Additionally, 8229 genes from an *E. coli* + *Shigella* dataset containing 42 genomes were studied for their correlated evolution with the capability of *E. coli* and *Shigella* spp. to infect human

(pathogenicity). The character states were reconstructed at ancestral nodes of the phylogenetic tree and character histories were traced for interesting pathways and enzymes that were correlated with the marine/non-marine living environment of *Rhodobacteraceae.* A gene correlated with the pathogenic capability of *E. coli* + *Shigella* was also reconstructed. Specific pathways, enzymes and genes were selected to infer different character history patterns over the phylogeny. The parsimony method was applied to reconstruct character states [Maddison, 1994] using the Mesquite software. This method traces the ancestral character states that minimize the number of steps of character change given the phylogenetic tree and character distribution data. There are two parsimony models available for categorical data in Mesquite software: ordered and unordered. The unordered states model was used to reconstruct character states. The ordered states model costs more than one step for a transformation from one state to another. For example, in the case of character transformation from state 1 to 3, the ordered states model costs 2 steps (1 to 2 and 2 to 3) where the unordered states model costs 1 step (1 to 3). As interval steps are not possible for presence/absence data, the ordered states model does not have an effect in character state reconstruction of presence/absence data. Multiple reconstructions of character evolution are possible for a single character on the given phylogeny. These are called the Most Parsimonious Reconstructions (MPRs). All of them are most parsimonious with the same number of steps and different character state transitions at specific nodes of the phylogeny. To summarize MPRs, the accelerated transformation criterion had been followed. The Mesquite software assigns uncertain character state transition at nodes where transitions differ between various MPRs [Maddison and Maddison, 2001]. The character history of the cp4-44 prophage element was comparatively traced along with the character history of pathogenicity in the *E. coli* + *Shigella* dataset on the phylogeny where 29 organisms are pathogens and 13 non-pathogens. The character state histories of four pathways and one enzyme were comparatively traced along with the character history of marine/non-marine living environments in the *Rhodobacteraceae* dataset where 88 organisms are marine and 20 non-marine.

4.4. RESULTS

4.4.1. THRESHOLD OPTIMIZATION

From the Pearson's chi-squared test described in the section 4.3.9, goodness of fit values for seven threshold values were plotted in box plot and shown in Figure 24.

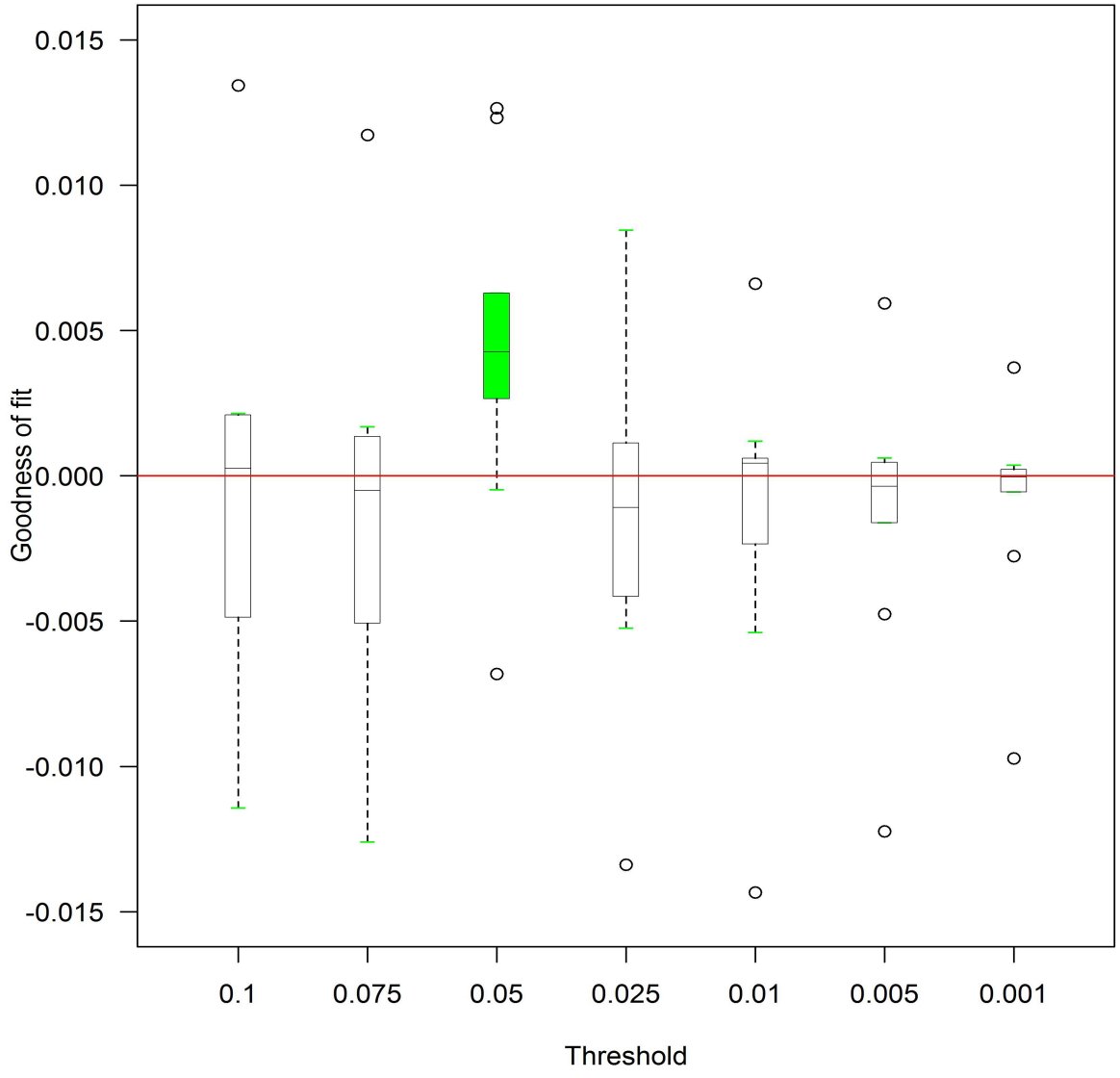


Figure 24. Boxplot – threshold vs goodness of fit. Each box represents the distribution of the goodness of fit values for different thresholds applied to p-values of the LR statistics of gene pairs from eight datasets. The green color highlights the threshold p<0.05. The red line separates the positive and

negative goodness of fit values.

Figure 24 shows the box plot of the applied threshold values *vs*. goodness of fit calculated by the chi-squared test for the fields 1 and 4 (example contingencies shown in Table 6) for eight datasets from Table S1. If the goodness of fit value is above zero, it indicates that observed number of gene pairs (which are related to each other by means of both correlated evolution and pathway involvement) are more than the expected frequency. Figure 22 shows a green shade in the third box where the threshold value was applied 0.05. In the green box, the upper and lower quartiles shows positive goodness of fit values. The minimum value and outlier show negative goodness of fit for the threshold 0.05. It means that the threshold value 0.05 is comparatively good enough to other threshold values to infer the gene pairs which evolved in correlated fashion and involved in same pathway. Thus, threshold value 0.05 is chosen to apply to p-values of LR statistics between likelihood of BayesTraits models with the assumptions of dependence/independence between gene pairs.

Although 0.01 and 0.001 yields more strict threshold values in terms of finding evolutionarily correlated gene pairs, the dependence between pathway involvement and evolutionary correlation were observed as best for the applied threshold 0.05. Moreover, 0.01 and 0.001 thresholds filter the major amount of gene pairs as insignificantly correlated. In this case, the study of evolutionary correlations becomes impossible for the several genes in the dataset.

4.4.2. EFFECT OF SMALL GENOMES

From the MCL clustering process, the *Planctomycetes* and *E. coli* datasets yielded 21 and 15 correlated gene clusters respectively. The datasets with small genomes were clustered as one big gene cluster. For example, the *Spirochaetae* dataset yielded one big cluster. This reduced number of correlated gene clusters occurred due to the effect of small genomes. A small genome represents a genome with the less number of genes compare to other genomes in the dataset.

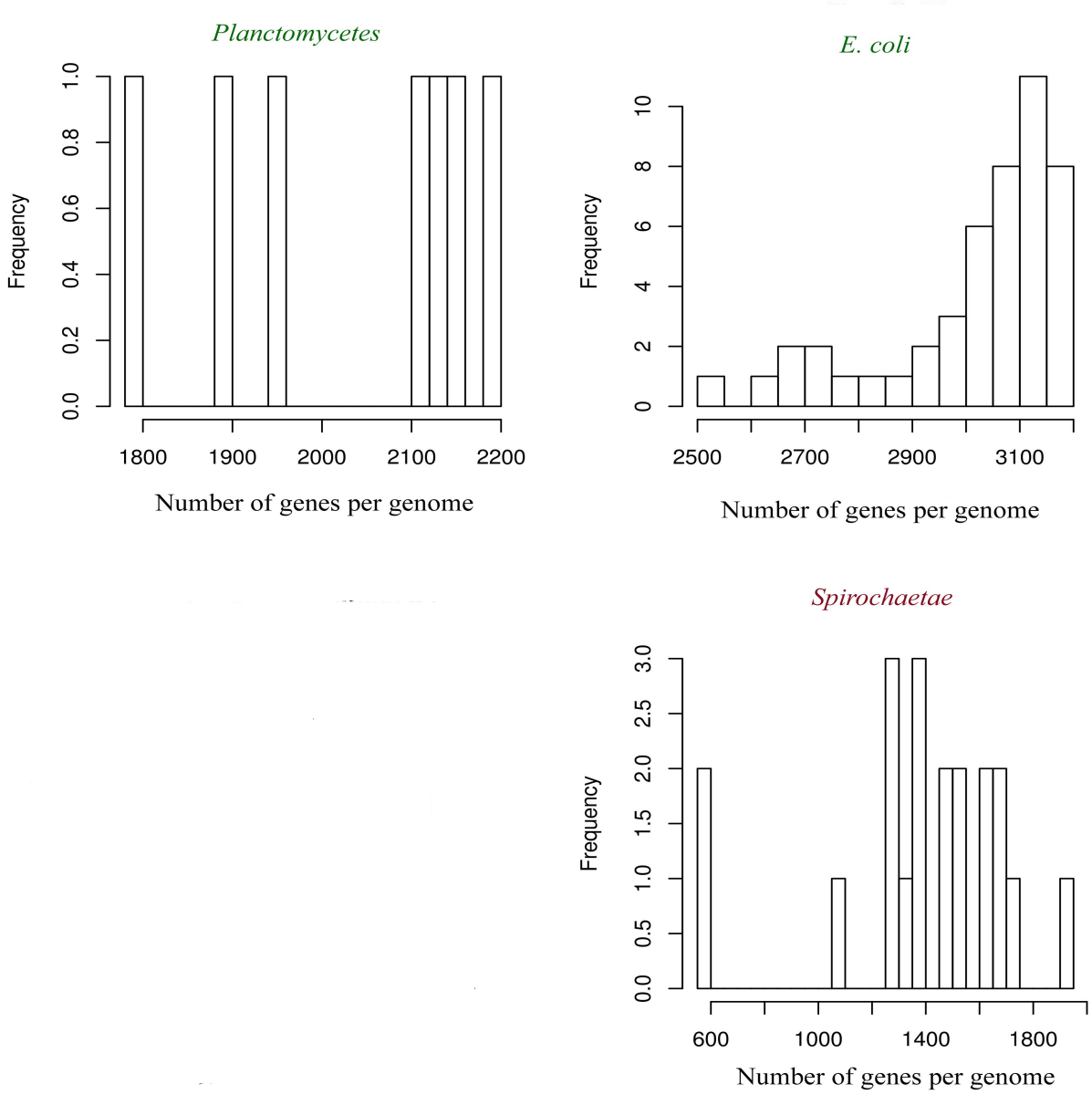


Figure 25. Distribution of genes per genome for Planctomycetes, E. coli and *Spirochaetae datasets. The green title indicates the distribution of gene counts per genome across the datasets. The red title indicates the distribution of gene counts per genome across datasets with large interval between small genomes and other genomes. The Spirochaetae dataset has small genomes with the gene counts less than 600 genes.*

In Figure 25, the gene counts per genome in *Planctomycetes* dataset shows distribution with the range 1750 to 2200 and the *E.coli* dataset as 2500 to 3200. The gene counts per genome in *Spirochaetae*

dataset shows distribution with the range 550 to 2000. In the case of the *Planctomycetes* and *E. coli* datasets, the distribution intervals between genes per genome are very close each other. In the case of the *Spirochaetae* dataset, the distribution intervals are very close each other from 1000 to 1900 and there is some outliers from 600 to 1100. The small genomes are *Borrelia burgdorferi* B31 and *Borrelia valaisiana* VS116. The *Spirocheatae* dataset was reduced without those two genomes and the number of correlated gene clusters increased to four from one. Further studies in *Spirochaetae* dataset proceeded by the remaining genomes without the above mentioned two genomes. The effect was caused due to the genes parallelly lost genes in the small genomes.

4.4.3. CORRELATED GENES IN FUNCTIONAL GROUPS

After the removal of small genomes, the MCL clustering process yielded 330 correlated gene clusters for the *Spirochaetae* dataset with 29 genomes. If a set of genes are clustered in a single cluster, it indicates that those genes are correlated evolutionarily with fellow genes in the same cluster.

COG groups and categories

The goodness of fit values for the COG groups versus correlated genes clusters using chi-squared test were obtained (described in the section 4.3.10). Some goodness of fit values are shown in Table 9. In Table 9, goodness of fit value 6.1857 is shown for the chi-squared contingency of correlated genes cluster numbers 213, 270, 319 and the "information storage and processing" COG group. It indicates that the observed number of genes is 6 times larger than the expected number of genes. Likewise, several positive goodness of fit values were observed for correlated genes clusters vs COG groups/categories. From the clusters of evolutionarily correlated and functionally linked genes, specific genes of interest can therefore be examined.

Table 9. Goodness of fit for COG groups - Spirochaetae dataset with 29 genomes

COG groups	Goodness of fit values	Cluster number
Information storage and processing	6.1857	213, 270, 319, 327, 328
Cellular processing and signaling	3.7581	280, 289, 294, 296
Metabolism	1.8778	221, 222, 231, 256, 265, 272
Poorly characterized	2.984	264, 273, 282, 298, 308, 315

Pathways

With the same procedure as above for the COG groups and categories, goodness of fit values were obtained from the chi-squared test (described in the section 4.3.10) for correlated gene clusters vs same pathway involvement. Several pathways with clusters of evolutionarily correlated genes were observed with goodness of fit value from Pearson's chi-squared test. Among those pathways, a study was further focused for motility pathways of the *Spirochaetae* dataset with 29 genomes. Motility pathways and genes involved in that pathway were especially considered in the *Spirochaetae* dataset because of the high interest and adequate research data available on the specific pathway with the experimental evidence [Abt et al., 2013]. Figure 26 shows the heatmap of the presence/absence of evolutionarily correlated genes involved in motility pathways of the genomes in the *Spirochaetae* dataset. The gene indices mentioned in the heat map can be mapped by Table S4.

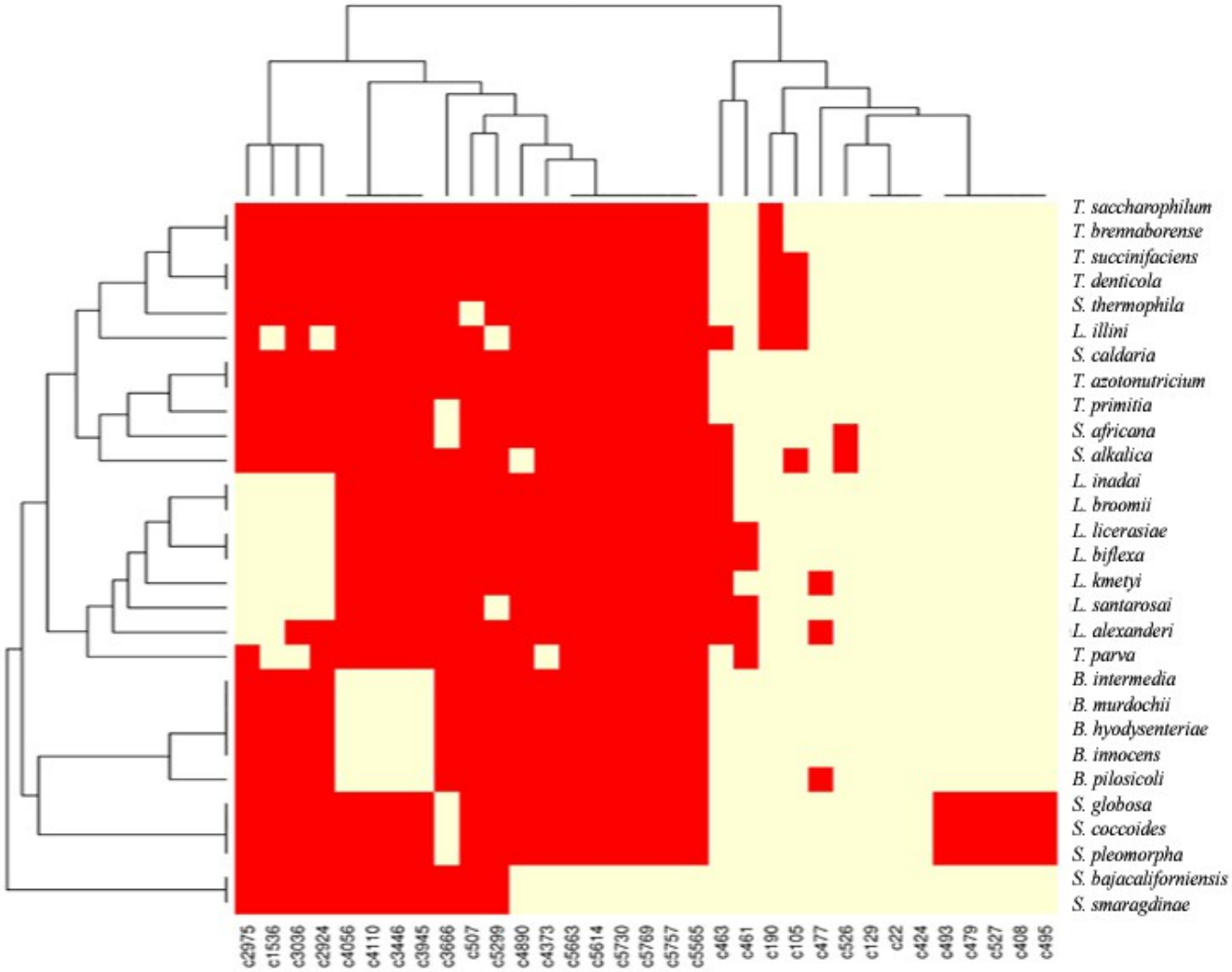


Figure 26. The presence/absence of correlated genes of motility pathways across 29 genomes of Spirochaetae dataset. The genes were plotted on the X axis and the genomes in the Spirochaetae dataset were plotted on the Y axis. The red color indicates the absence of genes and the white color indicates the presence of genes in the genome. The abbreviations for the x and y values are shown in Tbale S2 and S4.

Figure 26 infers: the non-motile organisms like *Sphaerochaeta globosa, Sphaerochaeta pleomorpha* and *Sphaerochaeta coccoides* [Droege et al., 2006, Ritalahti et al., 2012] are having the high number of absence in motility genes which are clustered as evolutionarily correlated. At the same time, highly motile *Spirochaeta bajacaliforniensi* [Fracek et al., 1985] has a high number of motility genes which are clustered as evolutionarily correlated.

4.4.4. DISTRIBUTION OF CORRELATED ENZYMES IN PATHWAYS

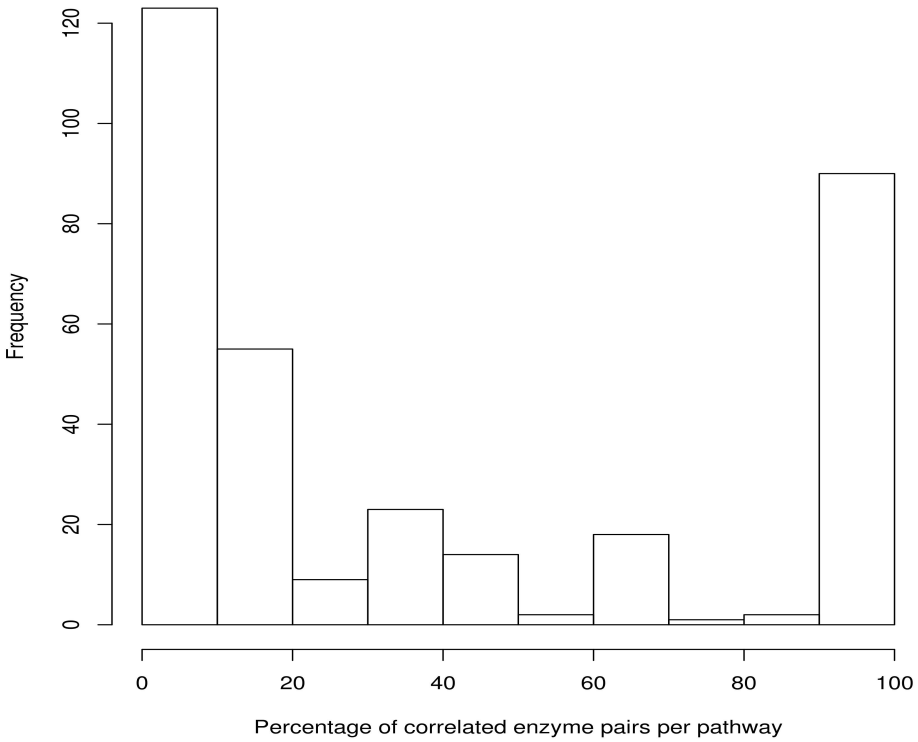


Figure 27 shows the distribution of the percentage of functionally linked enzymes involved in a pathway. For instance, if 10 enzymes involve in a pathway X and those 10 enzymes were observed in a single correlated gene cluster, it indicates that 100% of the enzymes in a pathway had evolved in a

correlated manner.*Figure 27. The distribution of correlated and functionally linked enzyme pairs in Rhodobacteraceae 87 dataset.*

Figure 27 indicates that there is a tendency in the histogram towards 100% and 0%. Of the 652 pathways, 90 pathways had enzymes which were correlated with each other in evolution. Some of the major pathways are completely encoded by enzymes which are evolutionarily correlated with each other. Some example pathways are: nitrate reduction I, beta-D-glucuronide and D-glucuronate degradation, D-galacturonate degradation I, dTDP-L-rhamnose biosynthesis I, catechol degradation to beta-ketoadipate and 5-dehydro-4-deoxy-D-glucuronate degradation. An example pathway is explained in the discussion.

4.4.5. CLUSTERS OF PATHWAYS

The MCL clustering of pathways yielded 11 clusters of evolutionarily correlated pathways. Some of the clustered pathways are 2,4,6-trichlorophenol degradation, 3,4,6-trichlorocatechol degradation and 1,4-dichlorobenzene degradation. The reason of the above clustered pathways are explained in the discussion section. The pathways present in all genomes were clustered as singlets in the MCL clustering. Those pathways were necessary metabolic pathways for survival of an organism.

4.4.6. CLUSTERS OF GENOMIC FEATURES

The 3 clusters along with 6 singlets of correlated genomic features of *Rhodobacteraceae* dataset were created based on the method described in the section 4.3.7. The clustered genomic features are shown in Table 10. The abbreviations for the genomic features are explained in Table S5.

Table 10. Clusters of genomic features - Rhodobacteraceae dataset with 87 genomes. The single letters represent the proportion of genes classified under COG categories.

Cluster number	Genomic features
1	Poorly characterized genes, R, S, SSU, LSU, genes involve in metabolism, E, tRNA, A, D, C, I, CRISPR, ribosomal proteins
2	Length of nucleotide sequence, number of genes, transporter proteins, ABC proteins, G, P, O, K, F, U, H, V, J
3	Information processing genes, L, genes involve in cellular processing, M, T, N, Q
4	GC content
5	B
6	Y
7	Z
8	W
9	Phage like elements

From Table 10, the rows four to eight were singlets from MCL clusters. The first three rows show the clustered genomic features which were correlated in evolution. This information help to understand the evolution of genomic elements. The evolutionary correlation of functionally related genomic elements can be observed using this strategy. An example from Table 10 are described in the discussion part.

4.4.7. EVOLUTIONARY CORRELATION OF GENES WITH PATHOGENESIS

In the dataset containing 8229 genes of the *E. coli* + *Shigella* dataset, 425 (4%) genes were identified as correlated in evolution with the pathogenicity. In particular, 57 genes were correlated in evolution with the pathogenicity and gained only in pathogens and 29 genes only in non-pathogens. The character history patterns of these genes are interesting as complete presence or absence was observed in a specific set of organisms (either pathogens or non-pathogens).

The cp4-44 prophage element was observed in only 14 pathogens of which nine are *E. coli* strains and five *Shigella* spp. The cp4-44 prophage was completely absent in non-pathogens of the dataset. The reconstructed character states of the cp4-44 prophage element at ancestral nodes of the phylogenetic tree are shown in Figure 28 as highly correlated (p-value = 0.0003) gene with pathogenicity.

It was observed in the left-hand tree of Figure 28 that the pathogenicity was inherited from the common

ancestor of the 42 organisms. One independent loss occurred at an ancestral node and independently in four extant organisms. At the ancestral node of *E. coli* strains BL21(DE3), B str. REL606, BW2952, K-12 str. DH10B, DH1, K-12 str. MG1655, K-12 str. W3110, ETEC H10407, and HS, the pathogenicity was lost and *E. coli* ETEC H10407 regained. A total of six losses and one regain of pathogenicity was observed on the phylogeny. In the right-hand tree of Figure 28, it was observed that the cp4-44 prophage element was not vertically inherited from the common ancestor of 42 organisms. It was independently gained on the phylogeny at ancestral nodes and the extant organisms where pathogenicity was already inherited. *Shigella boydii* strains and *Shigella sonnei* inherited the cp4-44 element from their ancestral node where the independent gain occurred. An uncertain reconstruction of a character state transition was observed at the ancestral node of *E. coli* UM146, *E. coli* UT189 and *E. coli* IHE3034. Due to uncertainty, the total number of character gains might be in the range range of eleven to twelve. The character loss might be one or no loss. Nine to eleven gains might have occurred independently in extant organisms with no ancestral inheritance. In a comparison between the two characters, it was observed that the pathogenicity and the cp4-44 phage element were correlated in evolution with different character history patterns on the phylogeny.

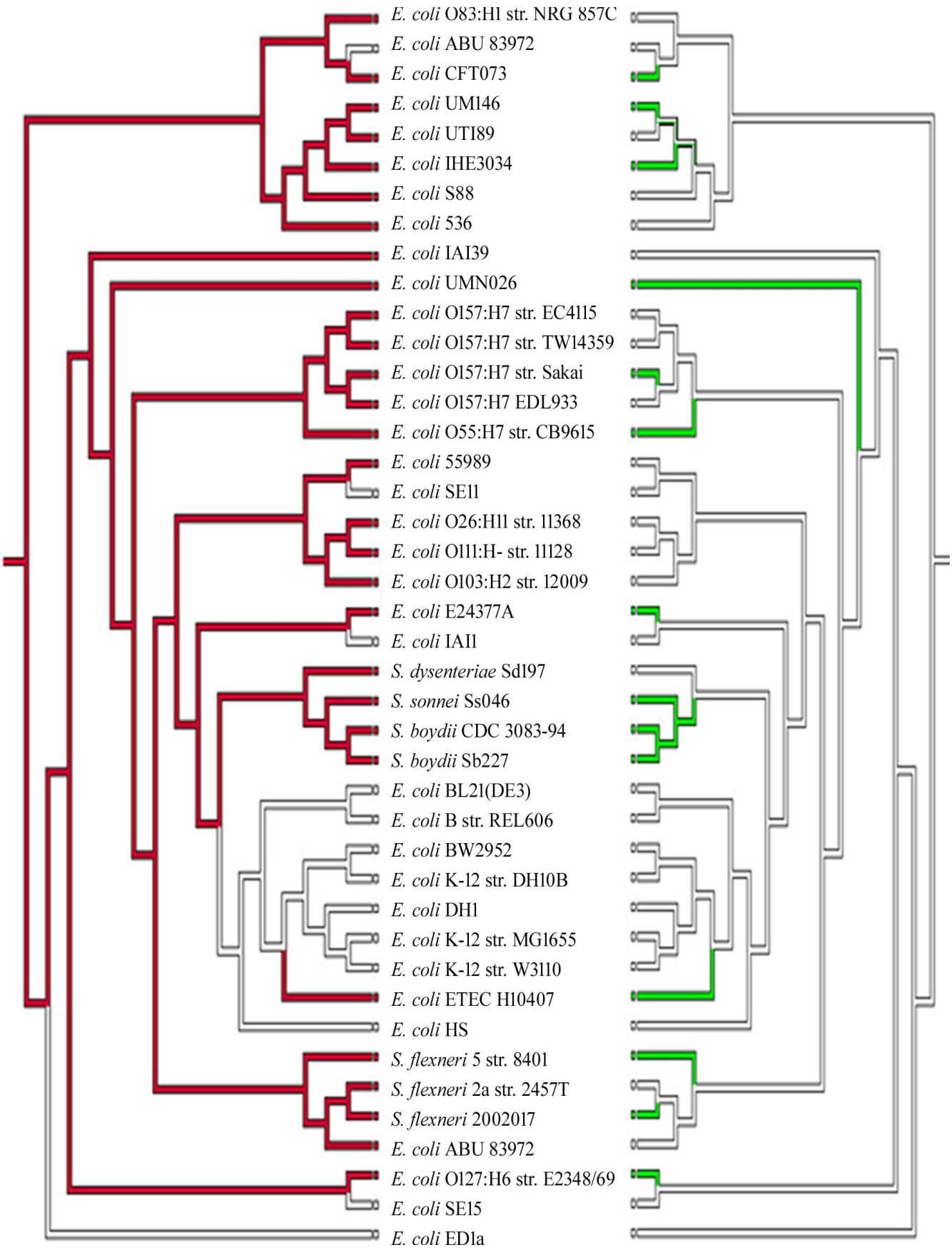

E. coli O83:H1 str. NRG 857C
E. coli ABU 83972
E. coli CFT073
E. coli UM146
E. coli UTI89
E. coli IHE3034
E. coli S88
E. coli 536
E. coli IAI39
E. coli UMN026
E. coli O157:H7 str. EC4115
E. coli O157:H7 str. TW14359
E. coli O157:H7 str. Sakai
E. coli O157:H7 EDL933
E. coli O55:H7 str. CB9615
E. coli 55989
E. coli SE11
E. coli O26:H11 str. 11368
E. coli O111:H- str. 11128
E. coli O103:H2 str. 12009
E. coli E24377A
E. coli IAI1
S. dysenteriae Sd197
S. sonnei Ss046
S. boydii CDC 3083-94
S. boydii Sb227
E. coli BL21(DE3)
E. coli B str. REL606
E. coli BW2952
E. coli K-12 str. DH10B
E. coli DH1
E. coli K-12 str. MG1655
E. coli K-12 str. W3110
E. coli ETEC H10407
E. coli HS
S. flexneri 5 str. 8401
S. flexneri 2a str. 2457T
S. flexneri 2002017
E. coli ABU 83972
E. coli O127:H6 str. E2348/69
E. coli SE15
E. coli ED1a

Figure 28. Both trees are the same phylogeny for 42 E. coli + Shigella strains. The left-hand tree shows the character history pattern of the pathogenicity. The red color indicates the presence of pathogenic character, white indicates absence. The right-hand tree shows the character history pattern of the cp4-44 prophage element. The green color indicates the presence of cp4-44 prophage element, white indicates absence. A color change from white to green and white to red at the nodes in the direction from the root to leaves on the phylogeny indicates the character state transition from the absence state to presence. The color change from green to whiite and red to white indicates the transition from the presence state to absence. The uncertain character state transitions are shown as dual color (green and white).

4.4.8. EVOLUTIONARY CORRELATION OF (S)-2-HALOACID DEHALOGENASE WITH LIVING ENVIRONMENTS OF RHODOBACTERACEAE

Among the enzymes of the *Rhodobacteraceae* dataset, 722 (39%) enzymes were identified as correlated in evolution with the living environments of *Rhodobacteraceae* spp. *(*either marine or non-marine). In particular, 73 enzymes were correlated in evolution with the living environment and observed only in marine organisms, and 45 enzymes only in non-marine organisms. The character history patterns of these enzymes are interesting as completely present or absent in either the marine or non-marine environment.

The (S)-2-haloacid dehalogenase was observed in 83 marine and four non-marine organisms of the dataset. The reconstructed character states of (S)-2-haloacid dehalogenase are shown on the right-hand phylogenetic tree in Figure 29 as an enzyme (p-value = 1.8×10^{-6}) highly correlated with the living environments of *Rhodobacteraceae*.

In the left-hand tree of Figure 29, it was observed that the 88 marine habitats were inherited in *Rhodobacteraceae* spp. from the common ancestor of 108 organisms. The loss of marine characteristics was observed at two ancestral nodes and independently in *Pannonibacter phragmitetus* DSM 14782. *Paracoccus zeaxanthinifaciens* ATCC 21588 and *Pseudorhodobacter ferrugineus* DSM 5888 are marine habitats independently regained the characteristics after a loss at the ancestral node of a clade containing 16 marine and 3 non-marine organisms. A total of three regains and losses were observed on the tree. In the right-hand tree, it was observed that the (S)-2-haloacid dehalogenase was inherited from the common ancestor of 108 *Rhodobacteraceae* on the phylogeny. It was independently lost at two

ancestral nodes and one extant organism. The organisms *Paracoccus denitrificans* PD1222, *Paracoccus denitrificans* SD1, *Paracoccus aminophilus* JCM7686 and *Pseudorhodobacter ferrugineus* DSM 5888 independently regained (S)-2-haloacid dehalogenase after a loss at the same ancestral node of 15 non-marine habitats over the tree. A total of four regains and three losses were observed on the phylogeny. In comparison between the two phylogenies with reconstructed characters, both characters were inherited from the common ancestor of 108 organisms and lost at the same ancestral nodes. After the loss at an ancestral node, regains occurred in different extant organisms for those two characters.

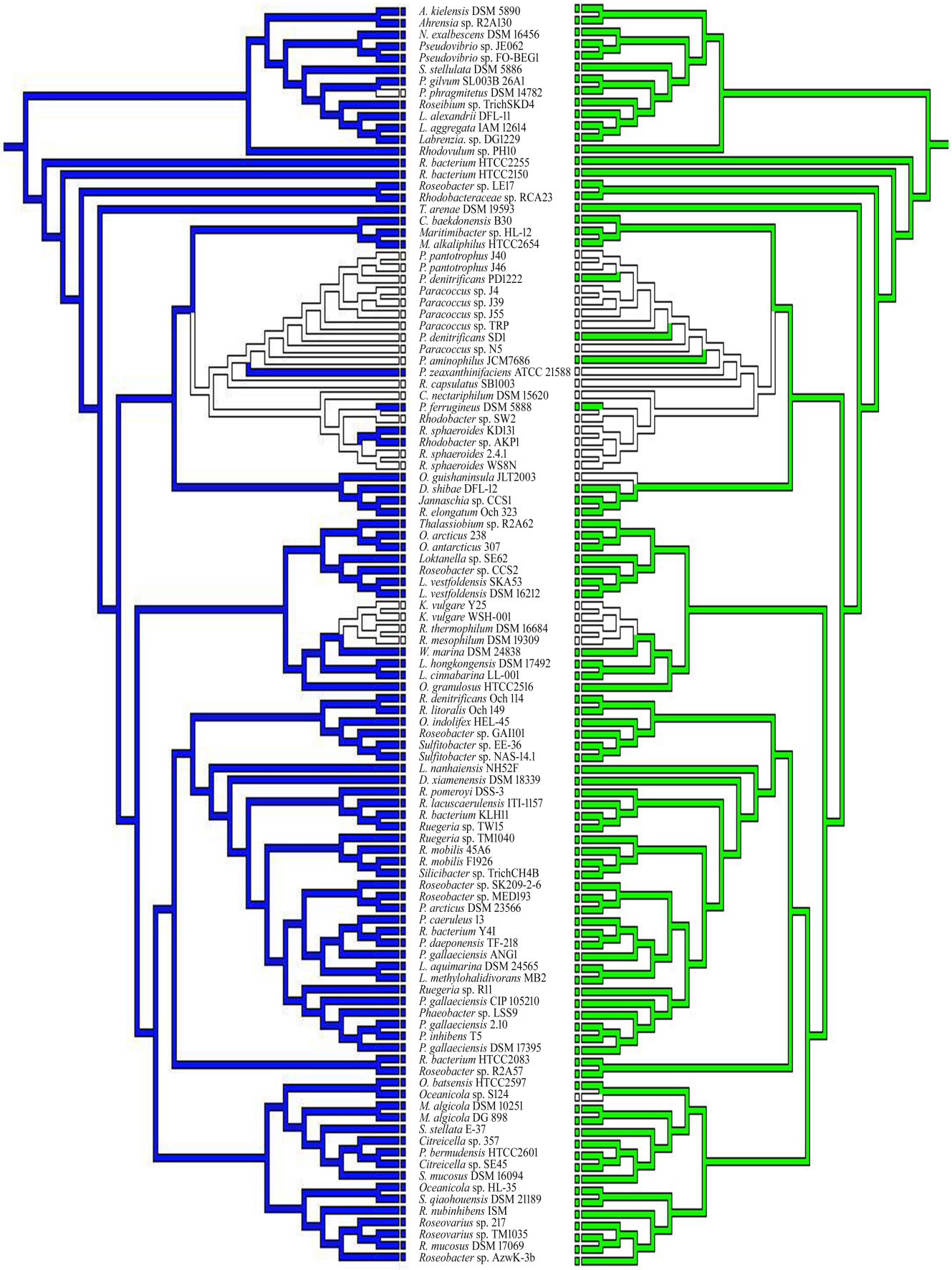

A. kielensis DSM 5890
Ahrensia sp. R2A130
N. exalbescens DSM 16456
Pseudovibrio sp. JE062
Pseudovibrio sp. FO-BEG1
S. stellulata DSM 5886
P. gilvum SL003B 26A1
P. phragmitetus DSM 14782
Roseibium sp. TrichSKD4
L. alexandrii DFL-11
L. aggregata IAM 12614
Labrenzia. sp. DG1229
Rhodovulum sp. PH10
R. bacterium HTCC2255
R. bacterium HTCC2150
Roseobacter sp. LE17
Rhodobacteraceae sp. RCA23
T. arenae DSM 19593
C. baekdonensis B30
Maritimibacter sp. HL-12
M. alkaliphilus HTCC2654
P. pantotrophus J40
P. pantotrophus J46
P. denitrificans PD1222
Paracoccus sp. J4
Paracoccus sp. J39
Paracoccus sp. J55
Paracoccus sp. TRP
P. denitrificans SD1
Paracoccus sp. N5
P. aminophilus JCM7686
P. zeaxanthinifaciens ATCC 21588
R. capsulatus SB1003
C. nectariphilum DSM 15620
P. ferrugineus DSM 5888
Rhodobacter sp. SW2
R. sphaeroides KD131
Rhodobacter sp. AKP1
R. sphaeroides 2.4.1
R. sphaeroides WS8N
O. guishaninsula JLT2003
D. shibae DFL-12
Jannaschia sp. CCS1
R. elongatum Och 323
Thalassiobium sp. R2A62
O. arcticus 238
O. antarcticus 307
Loktanella sp. SE62
Roseobacter sp. CCS2
L. vestfoldensis SKA53
L. vestfoldensis DSM 16212
K. vulgare Y25
K. vulgare WSH-001
R. thermophilum DSM 16684
R. mesophilum DSM 19309
W. marina DSM 24838
L. hongkongensis DSM 17492
L. cinnabarina LL-001
O. granulosus HTCC2516
R. denitrificans Och 114
R. litoralis Och 149
O. indolifex HEL-45
Roseobacter sp. GAI101
Sulfitobacter sp. EE-36
Sulfitobacter sp. NAS-14.1
L. nanhaiensis NH52F
D. xiamenensis DSM 18339
R. pomeroyi DSS-3
R. lacuscaerulensis ITI-1157
R. bacterium KLH11
Ruegeria sp. TW15
Ruegeria sp. TM1040
R. mobilis 45A6
R. mobilis F1926
Silicibacter sp. TrichCH4B
Roseobacter sp. SK209-2-6
Roseobacter sp. MED193
P. arcticus DSM 23566
P. caeruleus 13
R. bacterium Y4I
P. daeponensis TF-218
P. gallaeciensis ANG1
L. aquimarina DSM 24565
L. methylohalidivorans MB2
Ruegeria sp. R11
P. gallaeciensis CIP 105210
Phaeobacter sp. LSS9
P. gallaeciensis 2.10
P. inhibens T5
P. gallaeciensis DSM 17395
R. bacterium HTCC2083
Roseobacter sp. R2A57
O. batsensis HTCC2597
Oceanicola sp. S124
M. algicola DSM 10251
M. algicola DG 898
S. stellata E-37
Citreicella sp. 357
P. bermudensis HTCC2601
Citreicella sp. SE45
S. mucosus DSM 16094
Oceanicola sp. HL-35
S. qiaohouensis DSM 21189
R. nubinhibens ISM
Roseovarius sp. 217
Roseovarius sp. TM1035
R. mucosus DSM 17069
Roseobacter sp. AzwK-3b

Figure 29. Both trees are the same phylogeny for 108 Rhodobacteraceae spp. The left-hand tree shows the evolutionary character history pattern of the marine and non-marine habitats. The evolution of marine habitats were traced and marked in blue color. The right-hand tree shows the character history pattern of the (S)-2-haloacid dehalogenase. The green color indicates the presence of (S)-2-haloacid dehalogenase, white indicates absence. The color change from white to green and white to blue at the nodes in the direction from the root to leaves on the phylogeny indicates the character transition from the absence state to presence. The color change from green to white and blue to white indicates the transition from the presence state to absence. The uncertain character state transitions are shown as dual color (green and white).

4.4.9. EVOLUTIONARY CORRELATION OF PATHWAYS WITH LIVING ENVIRONMENTS OF RHODOBACTERACEAE

In the dataset containing 322 pathways, 159 (49%) pathways were identified as correlated in evolution with the living environment of *Rhodobacteraceae* spp. In particular, 15 pathways were correlated in evolution with the living environment and observed only in marine organisms. Nine pathways were observed only in non-marine organisms.

Ectoine biosynthesis pathway – observed only in marine organisms

The reconstructed character states of the ectoine biosynthesis pathway are shown on the right-hand phylogenetic tree in Figure 30 as a highly correlated (p-value = 0.005) pathway with the living environment as well as being present only in marine *Rhodobacteraceae.* It was observed in 34 marine organisms.

The observations in the left-hand tree are explained in section 4.4.8. In the right-hand tree of Figure 30, it was observed that the ectoine biosynthesis was gained in marine organisms later on the phylogeny and not inherited from the common ancestor of 108 organisms. The gains observed on the phylogeny were: an ancestral node on the phylogeny independently gained the pathway and inherited to 16 organisms over the phylogeny, an uncertainty in the character state reconstruction gain was observed at an ancestral node which could inherit the pathway to 11 organisms, an ancestral node on the phylogeny independently gained the pathway. It inherited to two organisms (*Maritimibacter* sp. HL-12 and *Maritimibacter alkaliphilus* HTCC2654) and five extant organisms independently gained the pathway. Due to the uncertainty, total pathway gains on the phylogeny might be in the range of eight to 13 and

the pathway loss was not observed. The extant organisms that gained ectoine biosynthesis pathways were *Nesiotobacter exalbescens* DSM 16456, *Stappia stellulata* DSM 5886, *Paracoccus zeaxanthinifaciens* ATCC 21588, *Oceanicola granulosus* HTCC2516, and *Oceanibulbus indolifex* HEL-45. Among 34 marine organisms containing ectoine biosynthesis pathway, four genera containing only one species (*Nesiotobacter exalbescens* DSM 16456, *Stappia stellulata* DSM 5886, *Silicibacter* sp. TrichCH4B, and *Stappia stellulata* DSM 5886) and one genus containing two species (*Maritimibacter* sp. HL-12 and *Maritimibacter alkaliphilus* HTCC2654) were observed. Figure 30 shows the dissimilarity that the ectoine biosynthesis pathway was independently gained from the root of the phylogeny and marine habitats vertically inherited from the root.

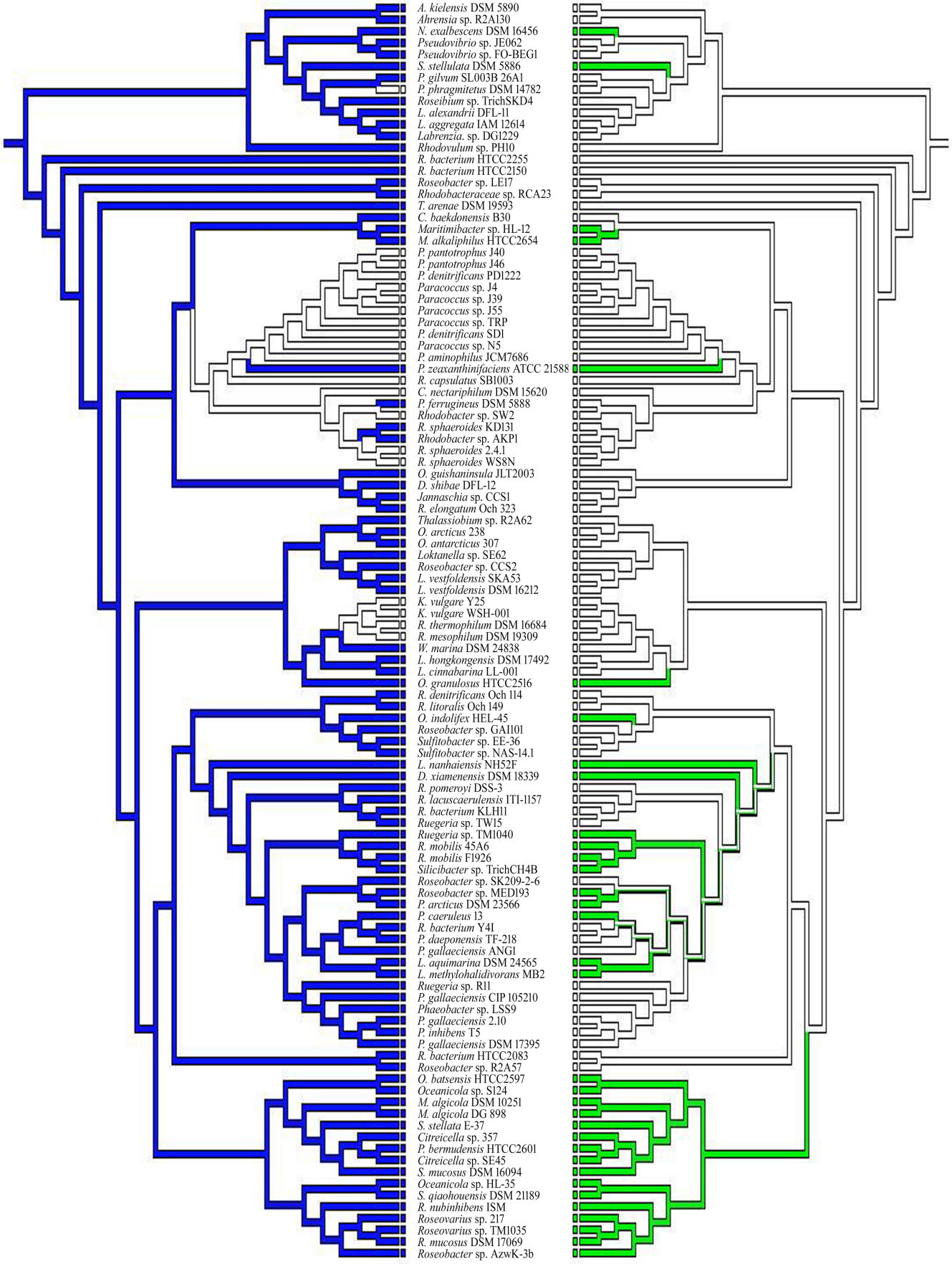

A. kielensis DSM 5890
Ahrensia sp. R2A130
N. exalbescens DSM 16456
Pseudovibrio sp. JE062
Pseudovibrio sp. FO-BEG1
S. stellulata DSM 5886
P. gilvum SL003B 26A1
P. phragmitetus DSM 14782
Roseibium sp. TrichSKD4
L. alexandrii DFL-11
L. aggregata IAM 12614
Labrenzia. sp. DG1229
Rhodovulum sp. PH10
R. bacterium HTCC2255
R. bacterium HTCC2150
Roseobacter sp. LE17
Rhodobacteraceae sp. RCA23
T. arenae DSM 19593
C. baekdonensis B30
Maritimibacter sp. HL-12
M. alkaliphilus HTCC2654
P. pantotrophus J40
P. pantotrophus J46
P. denitrificans PD1222
Paracoccus sp. J4
Paracoccus sp. J39
Paracoccus sp. J55
Paracoccus sp. TRP
P. denitrificans SD1
Paracoccus sp. N5
P. aminophilus JCM7686
P. zeaxanthinifaciens ATCC 21588
R. capsulatus SB1003
C. nectariphilum DSM 15620
P. ferrugineus DSM 5888
Rhodobacter sp. SW2
R. sphaeroides KD131
Rhodobacter sp. AKP1
R. sphaeroides 2.4.1
R. sphaeroides WS8N
O. guishaninsula JLT2003
D. shibae DFL-12
Jannaschia sp. CCS1
R. elongatum Och 323
Thalassiobium sp. R2A62
O. arcticus 238
O. antarcticus 307
Loktanella sp. SE62
Roseobacter sp. CCS2
L. vestfoldensis SKA53
L. vestfoldensis DSM 16212
K. vulgare Y25
K. vulgare WSH-001
R. thermophilum DSM 16684
R. mesophilum DSM 19309
W. marina DSM 24838
L. hongkongensis DSM 17492
L. cinnabarina LL-001
O. granulosus HTCC2516
R. denitrificans Och 114
R. litoralis Och 149
O. indolifex HEL-45
Roseobacter sp. GAI101
Sulfitobacter sp. EE-36
Sulfitobacter sp. NAS-14.1
L. nanhaiensis NH52F
D. xiamenensis DSM 18339
R. pomeroyi DSS-3
R. lacuscaerulensis ITI-1157
R. bacterium KLH11
Ruegeria sp. TW15
Ruegeria sp. TM1040
R. mobilis 45A6
R. mobilis F1926
Silicibacter sp. TrichCH4B
Roseobacter sp. SK209-2-6
Roseobacter sp. MED193
P. arcticus DSM 23566
P. caeruleus 13
R. bacterium Y4I
P. daeponensis TF-218
P. gallaeciensis ANG1
L. aquimarina DSM 24565
L. methylohalidivorans MB2
Ruegeria sp. R11
P. gallaeciensis CIP 105210
Phaeobacter sp. LSS9
P. gallaeciensis 2.10
P. inhibens T5
P. gallaeciensis DSM 17395
R. bacterium HTCC2083
Roseobacter sp. R2A57
O. batsensis HTCC2597
Oceanicola sp. S124
M. algicola DSM 10251
M. algicola DG 898
S. stellata E-37
Citreicella sp. 357
P. bermudensis HTCC2601
Citreicella sp. SE45
S. mucosus DSM 16094
Oceanicola sp. HL-35
S. qiaohouensis DSM 21189
R. nubinhibens ISM
Roseovarius sp. 217
Roseovarius sp. TM1035
R. mucosus DSM 17069
Roseobacter sp. AzwK-3b

Figure 30. Both trees are the same phylogeny for 108 Rhodobacteraceae spp. The left-hand tree shows the evolutionary character history pattern of the marine and non-marine habitats. The evolution of marine habitats were traced and marked in blue color. The right-hand tree shows the character history pattern of the ectoine biosynthesis pathway. The green color indicates the presence of ectoine biosynthesis pathway, white indicates absence. The color change from white to green and white to blue at the nodes in the direction from the root to leaves on the phylogeny indicates the character transition from the absence state to presence. The color change from green to white and blue to white indicates the transition from the presence state to absence. The uncertain character state transitions are shown as dual color (green and white).

2,3–dihydroxy benzoate biosynthesis pathway – observed only in non-marine organisms

The reconstructed character states of the 2,3-dihydroxy benzoate biosynthesis pathway are shown on the right-hand phylogenetic tree in Figure 31 as a (p-value = 0.02) pathway correlated with the living environment as well as being present only in eight non-marine *Rhodobacteraceae*.

The observations in left-hand tree are explained in section 4.4.8. In the right-hand tree of Figure 31, it was observed that the 2,3-dihydroxy benzoate biosynthesis pathway was uncertainly reconstructed at an ancestral node of nine non-marine organisms. Due to the uncertainty, the number of gains might be in the range of one to two with one or no loss. The organisms that gained the 2,3-dihydroxy benzoate biosynthesis pathway were *Paracoccus pantotrophus* J40, *Paracoccus pantotrophus* J46, *Paracoccus denitrificans* PD1222, *Paracoccus* sp. J4, *Paracoccus* sp. J39, *Paracoccus* sp. J55, *Paracoccus* sp. TRP and *Paracoccus* sp. N5. All organisms that contain the 2,3-dihydroxy benzoate biosynthesis pathway were *Paracoccus* spp. In comparison with 2,3-dihydroxy benzoate biosynthesis pathway and living environment of *Rhodobacteraceae* spp., the pathway was gained only in non-marine organisms on the phylogeny and correlated in evolution with the living environments of *Rhodobacteraceae* spp.

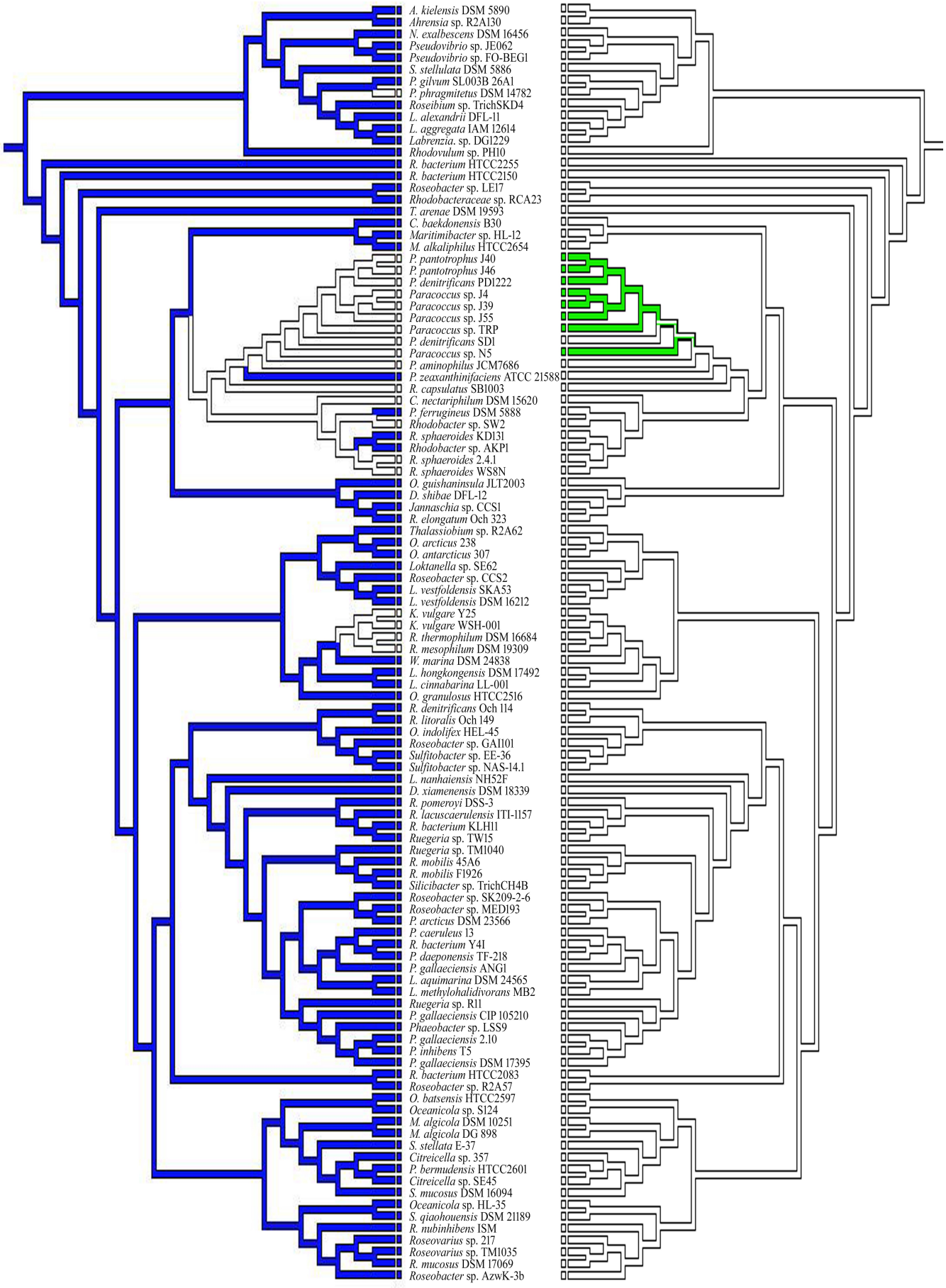

A. kielensis DSM 5890
Ahrensia sp. R2A130
N. exalbescens DSM 16456
Pseudovibrio sp. JE062
Pseudovibrio sp. FO-BEG1
S. stellulata DSM 5886
P. gilvum SL003B 26A1
P. phragmitetus DSM 14782
Roseibium sp. TrichSKD4
L. alexandrii DFL-11
L. aggregata IAM 12614
Labrenzia. sp. DG1229
Rhodovulum sp. PH10
R. bacterium HTCC2255
R. bacterium HTCC2150
Roseobacter sp. LE17
Rhodobacteraceae sp. RCA23
T. arenae DSM 19593
C. baekdonensis B30
Maritimibacter sp. HL-12
M. alkaliphilus HTCC2654
P. pantotrophus J40
P. pantotrophus J46
P. denitrificans PD1222
Paracoccus sp. J4
Paracoccus sp. J39
Paracoccus sp. J55
Paracoccus sp. TRP
P. denitrificans SD1
Paracoccus sp. N5
P. aminophilus JCM7686
P. zeaxanthinifaciens ATCC 21588
R. capsulatus SB1003
C. nectariphilum DSM 15620
P. ferrugineus DSM 5888
Rhodobacter sp. SW2
R. sphaeroides KD131
Rhodobacter sp. AKP1
R. sphaeroides 2.4.1
R. sphaeroides WS8N
O. guishaninsula JLT2003
D. shibae DFL-12
Jannaschia sp. CCS1
R. elongatum Och 323
Thalassiobium sp. R2A62
O. arcticus 238
O. antarcticus 307
Loktanella sp. SE62
Roseobacter sp. CCS2
L. vestfoldensis SKA53
L. vestfoldensis DSM 16212
K. vulgare Y25
K. vulgare WSH-001
R. thermophilum DSM 16684
R. mesophilum DSM 19309
W. marina DSM 24838
L. hongkongensis DSM 17492
L. cinnabarina LL-001
O. granulosus HTCC2516
R. denitrificans Och 114
R. litoralis Och 149
O. indolifex HEL-45
Roseobacter sp. GAI101
Sulfitobacter sp. EE-36
Sulfitobacter sp. NAS-14.1
L. nanhaiensis NH52F
D. xiamenensis DSM 18339
R. pomeroyi DSS-3
R. lacuscaerulensis ITI-1157
R. bacterium KLH11
Ruegeria sp. TW15
Ruegeria sp. TM1040
R. mobilis 45A6
R. mobilis F1926
Silicibacter sp. TrichCH4B
Roseobacter sp. SK209-2-6
Roseobacter sp. MED193
P. arcticus DSM 23566
P. caeruleus 13
R. bacterium Y4I
P. daeponensis TF-218
P. gallaeciensis ANG1
L. aquimarina DSM 24565
L. methylohalidivorans MB2
Ruegeria sp. R11
P. gallaeciensis CIP 105210
Phaeobacter sp. LSS9
P. gallaeciensis 2.10
P. inhibens T5
P. gallaeciensis DSM 17395
R. bacterium HTCC2083
Roseobacter sp. R2A57
O. batsensis HTCC2597
Oceanicola sp. S124
M. algicola DSM 10251
M. algicola DG 898
S. stellata E-37
Citreicella sp. 357
P. bermudensis HTCC2601
Citreicella sp. SE45
S. mucosus DSM 16094
Oceanicola sp. HL-35
S. qiaohouensis DSM 21189
R. nubinhibens ISM
Roseovarius sp. 217
Roseovarius sp. TM1035
R. mucosus DSM 17069
Roseobacter sp. AzwK-3b

Figure 31. Both trees are the same phylogeny for 108 Rhodobacteraceae spp. The left-hand tree shows the evolutionary character history pattern of the marine and non-marine habitats. The evolution of marine habitats were traced and marked in blue color. The right-hand tree shows the character history pattern of the 2,3-dihydroxy benzoate biosynthesis pathway. The green color indicates the presence of 2,3-dihydroxy benzoate biosynthesis pathway, white indicates absence. The color change from white to green and white to blue at the nodes in the direction from the root to leaves on the phylogeny indicates the character transition from the absence state to presence. The color change from green to white and blue to white indicates the transition from the presence state to absence. The uncertain character state transitions are shown as dual color (green and white).

Ethanol degradation pathway IV – character history pattern with the distributed presence of a pathway among both marine and non-marine organisms

The reconstructed character states of ethanol degradation pathway IV are shown on the right-hand phylogenetic tree in Figure 32 as pathway correlated with living environment (p-value = 0.02) and were observed in 17 out of 20 (85%) non-marine organisms and 43 out of 88 (49%) marine organisms.

The observations from the left-hand tree are explained in section 4.4.8. In the right-hand tree of Figure 32, uncertain inheritance of the ethanol degradation pathway IV from the common ancestor of 108 organisms was observed on the phylogeny.. Further, the total number of ethanol degradation pathway IV gain/regain might be in the range of eight to 26 with losses in the range of four to 18 on the phylogeny. Irrespective of uncertainty, five extant organisms were observed with the independent gain of the ethanol degradation pathway IV on the given phylogeny. They were *Ruegeria pomeroyi* DSS-3, *Ruegeria* sp. TW15, *Ruegeria* mobilis F1926, *Roseobacter* sp. SK209-2-6, and *Roseovarius* sp. TM1035. The comparative observation between two characters in Figure 32 shows the dissimilar character history patterns where ethanol degradation pathway IV has several uncertain gains and losses on the phylogeny than the characteristics of living in marine environment.

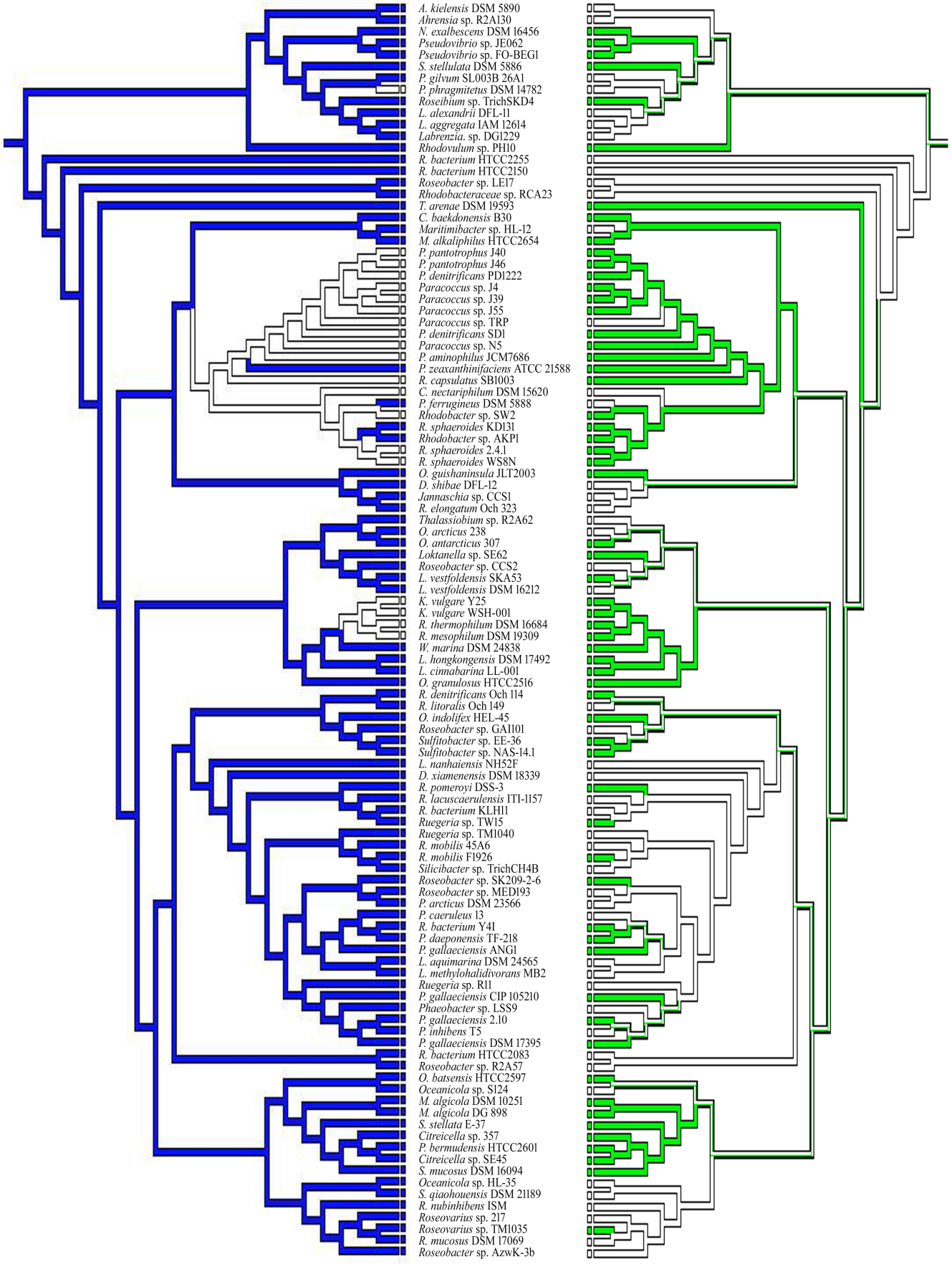

A. kielensis DSM 5890
Ahrensia sp. R2A130
N. exalbescens DSM 16456
Pseudovibrio sp. JE062
Pseudovibrio sp. FO-BEG1
S. stellulata DSM 5886
P. gilvum SL003B 26A1
P. phragmitetus DSM 14782
Roseibium sp. TrichSKD4
L. alexandrii DFL-11
L. aggregata IAM 12614
Labrenzia. sp. DG1229
Rhodovulum sp. PH10
R. bacterium HTCC2255
R. bacterium HTCC2150
Roseobacter sp. LE17
Rhodobacteraceae sp. RCA23
T. arenae DSM 19593
C. baekdonensis B30
Maritimibacter sp. HL-12
M. alkaliphilus HTCC2654
P. pantotrophus J40
P. pantotrophus J46
P. denitrificans PD1222
Paracoccus sp. J4
Paracoccus sp. J39
Paracoccus sp. J55
Paracoccus sp. TRP
P. denitrificans SD1
Paracoccus sp. N5
P. aminophilus JCM7686
P. zeaxanthinifaciens ATCC 21588
R. capsulatus SB1003
C. nectariphilum DSM 15620
P. ferrugineus DSM 5888
Rhodobacter sp. SW2
R. sphaeroides KD131
Rhodobacter sp. AKP1
R. sphaeroides 2.4.1
R. sphaeroides WS8N
O. guishaninsula JLT2003
D. shibae DFL-12
Jannaschia sp. CCS1
R. elongatum Och 323
Thalassiobium sp. R2A62
O. arcticus 238
O. antarcticus 307
Loktanella sp. SE62
Roseobacter sp. CCS2
L. vestfoldensis SKA53
L. vestfoldensis DSM 16212
K. vulgare Y25
K. vulgare WSH-001
R. thermophilum DSM 16684
R. mesophilum DSM 19309
W. marina DSM 24838
L. hongkongensis DSM 17492
L. cinnabarina LL-001
O. granulosus HTCC2516
R. denitrificans Och 114
R. litoralis Och 149
O. indolifex HEL-45
Roseobacter sp. GAI101
Sulfitobacter sp. EE-36
Sulfitobacter sp. NAS-14.1
L. nanhaiensis NH52F
D. xiamenensis DSM 18339
R. pomeroyi DSS-3
R. lacuscaerulensis ITI-1157
R. bacterium KLH11
Ruegeria sp. TW15
Ruegeria sp. TM1040
R. mobilis 45A6
R. mobilis F1926
Silicibacter sp. TrichCH4B
Roseobacter sp. SK209-2-6
Roseobacter sp. MED193
P. arcticus DSM 23566
P. caeruleus 13
R. bacterium Y4I
P. daeponensis TF-218
P. gallaeciensis ANG1
L. aquimarina DSM 24565
L. methylohalidivorans MB2
Ruegeria sp. R11
P. gallaeciensis CIP 105210
Phaeobacter sp. LSS9
P. gallaeciensis 2.10
P. inhibens T5
P. gallaeciensis DSM 17395
R. bacterium HTCC2083
Roseobacter sp. R2A57
O. batsensis HTCC2597
Oceanicola sp. S124
M. algicola DSM 10251
M. algicola DG 898
S. stellata E-37
Citreicella sp. 357
P. bermudensis HTCC2601
Citreicella sp. SE45
S. mucosus DSM 16094
Oceanicola sp. HL-35
S. qiaohouensis DSM 21189
R. nubinhibens ISM
Roseovarius sp. 217
Roseovarius sp. TM1035
R. mucosus DSM 17069
Roseobacter sp. AzwK-3b

Figure 32. Both trees are the same phylogeny for 108 Rhodobacteraceae spp. The left-hand tree shows the evolutionary character history pattern of the marine and non-marine habitats. The evolution of marine habitats were traced and marked in blue color. The right-hand tree shows the character history pattern of the ethanol degradation pathway IV. The green color indicates the presence of ethanol degradation pathway IV, white indicates absence. The color change from white to green and white to blue at the nodes in the direction from the root to leaves on the phylogeny indicates the character transition from the absence state to presence. The color change from green to white and blue to white indicates the transition from the presence state to absence. The uncertain character state transitions are shown as dual color (green and white).

Glycogen biosynthesis pathway I – uncorrelated in evolution with living environments of *Rhodobacteraceae* spp.

The observations in the left-hand tree are explained in section 4.4.8. The character history of glycogen biosynthesis pathway I is shown on the right-hand phylogenetic tree in Figure 33. The glycogen biosynthesis pathway I was observed in 16 out of 20 (80%) non-marine organisms and 30 out of 88 (34%) marine organisms. In the character history pattern of glycogen pathway I, it was observed that one uncertain gain occurred at an ancestral node and three extant organisms independently gained the pathway without any ancestral inheritance on the given phylogeny. Due to an uncertainty, the number of pathway gains might be in the range of six to seven with six losses on the phylogeny. The pathway was not inherited from the common ancestor of 108 organisms. By comparatively studying two characters in Figure 33, it was observed that glycogen biosynthesis pathway I was not inherited from the root of *Rhodobacteraceae* spp. where marine habitats were inherited. The evolution of glycogen biosynthesis pathway I was also uncorrelated with the evolution of marine habitats.

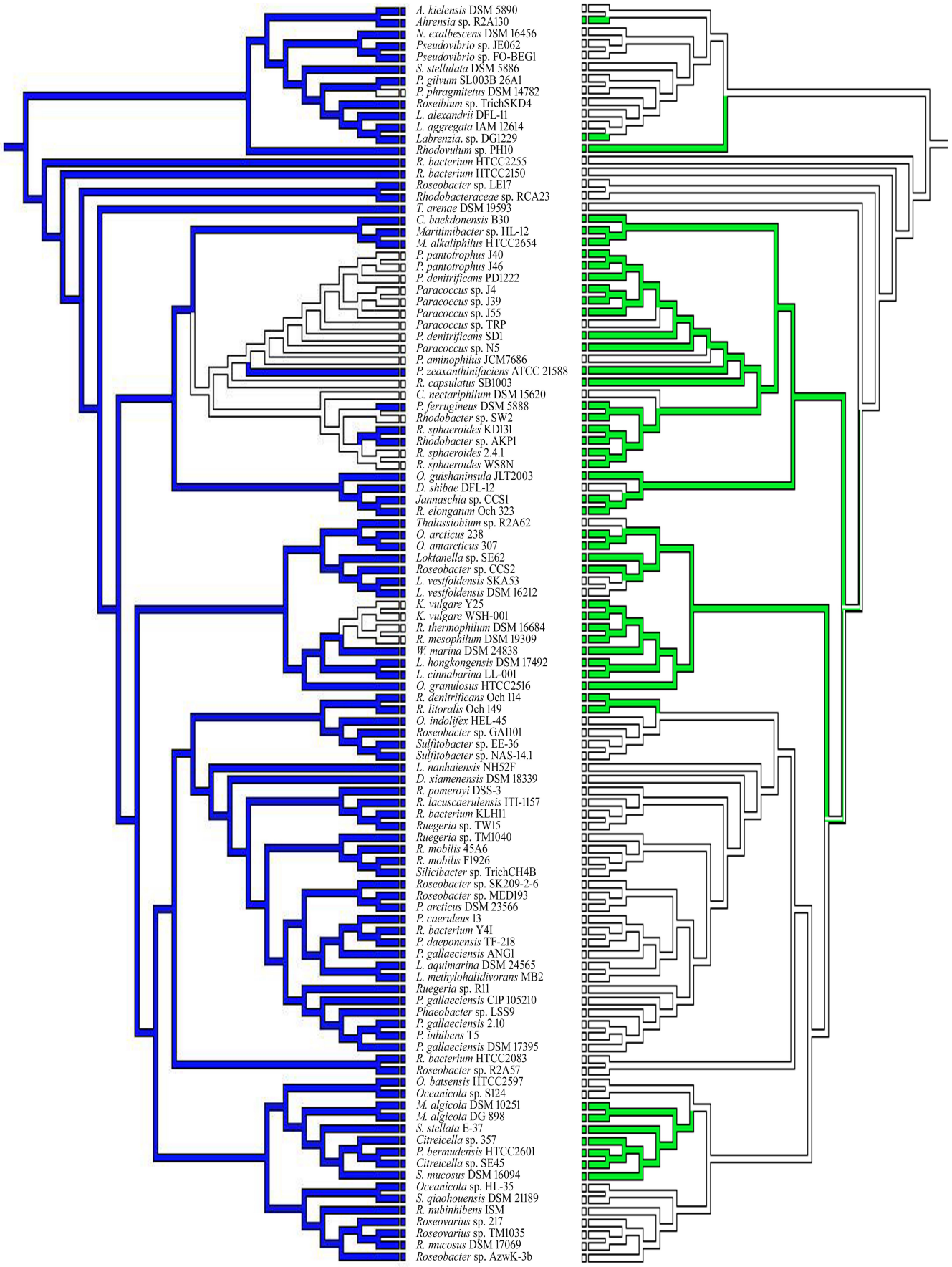

A. kielensis DSM 5890
Ahrensia sp. R2A130
N. exalbescens DSM 16456
Pseudovibrio sp. JE062
Pseudovibrio sp. FO-BEG1
S. stellulata DSM 5886
P. gilvum SL003B 26A1
P. phragmitetus DSM 14782
Roseibium sp. TrichSKD4
L. alexandrii DFL-11
L. aggregata IAM 12614
Labrenzia. sp. DG1229
Rhodovulum sp. PH10
R. bacterium HTCC2255
R. bacterium HTCC2150
Roseobacter sp. LE17
Rhodobacteraceae sp. RCA23
T. arenae DSM 19593
C. baekdonensis B30
Maritimibacter sp. HL-12
M. alkaliphilus HTCC2654
P. pantotrophus J40
P. pantotrophus J46
P. denitrificans PD1222
Paracoccus sp. J4
Paracoccus sp. J39
Paracoccus sp. J55
Paracoccus sp. TRP
P. denitrificans SD1
Paracoccus sp. N5
P. aminophilus JCM7686
P. zeaxanthinifaciens ATCC 21588
R. capsulatus SB1003
C. nectariphilum DSM 15620
P. ferrugineus DSM 5888
Rhodobacter sp. SW2
R. sphaeroides KD131
Rhodobacter sp. AKP1
R. sphaeroides 2.4.1
R. sphaeroides WS8N
O. guishaninsula JLT2003
D. shibae DFL-12
Jannaschia sp. CCS1
R. elongatum Och 323
Thalassiobium sp. R2A62
O. arcticus 238
O. antarcticus 307
Loktanella sp. SE62
Roseobacter sp. CCS2
L. vestfoldensis SKA53
L. vestfoldensis DSM 16212
K. vulgare Y25
K. vulgare WSH-001
R. thermophilum DSM 16684
R. mesophilum DSM 19309
W. marina DSM 24838
L. hongkongensis DSM 17492
L. cinnabarina LL-001
O. granulosus HTCC2516
R. denitrificans Och 114
R. litoralis Och 149
O. indolifex HEL-45
Roseobacter sp. GAI101
Sulfitobacter sp. EE-36
Sulfitobacter sp. NAS-14.1
L. nanhaiensis NH52F
D. xiamenensis DSM 18339
R. pomeroyi DSS-3
R. lacuscaerulensis ITI-1157
R. bacterium KLH11
Ruegeria sp. TW15
Ruegeria sp. TM1040
R. mobilis 45A6
R. mobilis F1926
Silicibacter sp. TrichCH4B
Roseobacter sp. SK209-2-6
Roseobacter sp. MED193
P. arcticus DSM 23566
P. caeruleus 13
R. bacterium Y4I
P. daeponensis TF-218
P. gallaeciensis ANG1
L. aquimarina DSM 24565
L. methylohalidivorans MB2
Ruegeria sp. R11
P. gallaeciensis CIP 105210
Phaeobacter sp. LSS9
P. gallaeciensis 2.10
P. inhibens T5
P. gallaeciensis DSM 17395
R. bacterium HTCC2083
Roseobacter sp. R2A57
O. batsensis HTCC2597
Oceanicola sp. S124
M. algicola DSM 10251
M. algicola DG 898
S. stellata E-37
Citreicella sp. 357
P. bermudensis HTCC2601
Citreicella sp. SE45
S. mucosus DSM 16094
Oceanicola sp. HL-35
S. qiaohouensis DSM 21189
R. nubinhibens ISM
Roseovarius sp. 217
Roseovarius sp. TM1035
R. mucosus DSM 17069
Roseobacter sp. AzwK-3b

Figure 33. Both trees are the same phylogeny for 108 Rhodobacteraceae spp. The left-hand tree shows the evolutionary character history pattern of the marine and non-marine habitats. The evolution of marine habitats were traced and marked in blue color. The right-hand tree shows the character history pattern of the glycogen biosynthesis pathway I. The green color indicates the presence of glycogen biosynthesis pathway I, white indicates absence. The color change from white to green and white to blue at the nodes in the direction from the root to leaves on the phylogeny indicates the character transition from the absence state to presence. The color change from green to white and blue to white indicates the transition from the presence state to absence. The uncertain character state transitions are shown as dual color (green and white).

Overview of the *Rhodobacteraceae* dataset in correlation study In the case of the *Rhodobacteraceae* dataset, the percentage of characters correlated with the living environment increased from low level functional features (gene-content (17%), ortholog-content (19%)) to high level (enzymes (39%) and pathways (49%)). The list of pathways correlated with the living environment of *Rhodobacteraceae* spp. were shown in Table S6. The number of correlated enzymes, genes, and orthologs correlated with living environment are more and they can not be listed in the thesis.

4.5. DISCUSSION AND CONCLUSION

A strategy for finding evolutionary correlations between functionally linked characters was established for genes, enzymes, pathways and genomic features and standardized using different datasets. Selected results were verified with respect to the references.

Threshold optimization

In the case of large sample size, the selection of a threshold is superior to choose an arbitrary alpha value [Barbash et al., 2013]. The best alpha value varies based on the characteristics of data. Thus, eight different datasets with 7 to 54 genomes, respectively, were used to optimize the threshold value. Thus threshold value for identifying significantly correlated gene pairs were optimized and result were shown in the section 4.4.1 as 0.05 threshold. The chi-squared model applied with the threshold value 0.05 was chosen by neglecting the other models which contribute non-significantly to improve the goodness of fit. Still, the 0.05 threshold value is suitable only for p-values obtained from LR test between likelihoods of BayesTrait's dependence/independence model of discrete characters (genes/enzymes/pathways). It has to be optimized for different types of data. In the case of genomic features, 0.01 was chosen as threshold value to get the significantly correlated genomic features. Because genomic features were highly correlated to each other when compared to discrete characters. Thus, p-values obtained from LR test between likelihoods of BayesTrait's continuous character (genomic features) models need more strict threshold value than discrete characters.

Effect of small genomes

Most of those genes in small genomes are basic genes needed for the survival of an organism. Those small genomes lost several genes which are present in other genomes in a parallel manner. If the small sized genomes are included in the correlated evolutionary analysis of genes across several genomes with large number of genes, the loss of genes in the small genomes causes the genes to form a single big clusters of correlated genes. Because, the parallel loss of several genes in certain genomes causes the BayesTraits software models to show same evolutionary correlation with all other genes in the dataset. Thus, the clustering of genes with respect to the evolutionary correlations were affected. In the case of the *Spirochaetae* dataset, small genomes (*Borrelia* spp) were identified from the distribution of genes per genome from the results shown in the section 4.4.2. The small genomes (*Borrelia* spp) in the *Spirochaetae* dataset were parasite genomes [Barbour and Stanley, 1986] with the loss of genes under

several COG categories [Abt et al., 2013]. The increase in the number of clusters in the clustering process after removal of *Borrelia* species and evidence of loss of genes under several COG categories in the same *Borrelia* species shows a relative confirmation that small genomes with lost genes effect the clustering process of correlated gene pairs. Theoretically, the same effect may be observed in the case of parallel genes gained among a specific group of organisms in the dataset. It is concluded that the parallel loss/gain of genes in a group or organisms in the specific dataset should be removed to cluster correlated genes using the implemented strategy.

Correlated gene pairs versus genes observed from same COG group/categories and pathways

The inference of correlated gene pairs in single COG group/category indicates that those genes were evolved together for the same need. This information can be further used to understand the gain of functions in an organisms with respect to specific function encoded genes. For example, genes under cell motility (an example COG category). The gain of motile characteristics is dependent to the correlated evolution of motility genes. The correlated evolution of genes in the same pathway is a hint to understand the pathway evolution and networking of pathway. In this perspective, specific motility pathways in the *Spirochaetae* dataset were studied and an evolutionary correlation of genes involved in motility pathways was cross verified with regard to the previous experimental evidence.

According to the study of [Abt et al., 2013], the number of genes responsible for flagellation are directly proportional to the number of flagella in *Spirochaetae* spp. It indicates the evolutionary correlation between motility genes. According to studies in [Droege et al., 2006], *Sphaerochaeta coccoides* is non-motile, helical in shape and no axial filament observations. The study in [Ritalahti et al., 2012] reports flagella-like formations in the periplasmic space of the *Sphaerochaeta pleomorpha* cell; furthermore, the organism does not have complete flagella for motility. The *Sphaerochaeta globosa* is also reported as a non-motile organisms. From the two studies stated above, it is clear that *Sphaerochaeta pleomorpha* and *Sphaerochaeta coccoides* are not bacteria with multiflagellar formations, and they are not motile due to flagella. The heat map in Figure 26 indicates that *Sphaerochaeta pleomorpha, Sphaerochaeta globosa* and *Sphaerochaeta coccoides* have fewer evolutionarily correlated genes responsible for the motility. The highly motile *S. bajacaliforniensi* [Fracek et al., 1985] shows high number of correlated genes in motility pathways. Thus, the result observed from Figure 26 makes sense with respect to the previous studies [Abt et al., 2013], [Droege et al., 2006] and [Ritalahti et al., 2012]. The motility pathway is an example interpreted with respect to

the references for evolutionary correlation studies of functionally linked genes. This evolutionary correlation study of genes involved in same pathway can be further proceeded for various pathways and datasets.

Correlated enzyme pairs versus their pathway involvement

As the *Rhodobacteraceae* lineage is for particular interest of hydrocarbon degradation [Buchan and Gonzalez, 2010], the result in the section 4.4.4 highlights the hydrocarbon degradation pathways encoded completely by evolutionarily correlated enzymes. The hint towards hydrocarbon degradation pathways with completely correlated enzymes would be useful in exploring the underlying biochemical mechanism of those pathways. From the above examples, the 5-dehydro-4-deoxy-D-glucuronate degradation pathway is also called pectin degradation pathway. Pectin is a heteropolymer compound normally observed in the cell wall of plants. The most of enzymes in this pathway are controlled by the transcriptional regulator *kdgR* [Rodionov et al., 2004]. The pectin degradation pathway was identified in *Rhodobacteraceae* dataset and encoded completely by enzymes evolved in correlated fashion.

From both kinds of evidence, it is hypothesized that pectin degradation is gained with respect to the enzyme evolution in *Rhodobacteraceae* spp. with respect to the correlated evolution of its enzymes. Likewise, along with the regulator evidence of a pathway and the evolutionary correlation evidence of its enzymes, the evolution of a single pathway or possibilities of pathway extensions can be inferred. The hypotheses derived from this approach can be used to narrow down the focus towards group of taxa in *Rhodobacteraceae* dataset which are encoding a specific pathway.

Evolutionary correlation between pathways

The chloroaromatic hydrocarbon degradation pathways which are correlated in evolution were shown in section 4.4.5. All the above pathways are connected in a metabolic network of xenobiotic degradation of chlorobenzene compound. Thus, a part of the chlorobenzene degradation system in the *Rhodobacteraceae* lineage was evolved together. Chlorobenzene is used in the manufacturing process of pesticides [Bailey, 2001].

The significant evolutionary correlation between two pathways provides an insight into those pathways which are evolutionarily evolved together. This provides the opportunity to reconstruct a molecular network based on pathway adaptations and evolution in order to resolve the complications in molecular activities of an organism. The correlated evolution of pathways which involve in same metabolic network can be identified using this approach like an above example.

Evolutionary correlation between genomic features

The clustered genomic features in a group of genomes from the *Rhodobacteraceae* dataset provide information on dependencies amongst the genomic features. This facilitates the understanding of the molecular construction of genomes and their evolution. The overview of clusters were shown in Table 10. For example, the structural proteins were clustered as singlets and functionally linked proteins were clustered together. The functionally linked genomic features like proportion of genes involved in cell wall biogenesis (M), cell motility (N) and cell signaling were clustered together. All those genes were involving in cellular processes and signaling. Likewise, the evolutionary correlation and functional links between specific genomic features can be further explored using the clusters shown. This study can be applied for more genomic features like number of transposons and satellite DNA sequences to find the evolutionary correlation with other genomic elements and trace functional links between them.

Evolutionary correlation of cp4-44 prophage element with pathogenicity

Prophages are genetic elements gained from bacteriophages in bacteria. Among prophage elements which encode virulent genes in pathogenic bacteria, one of the genetic elements gained from bacteriophages is cp4-44 prophage [Brüssow et al., 2004]. It includes 7 genes (*yoeA*, *yoeG*, *yeeW*, *yoeF, yoeH*, *yoeD* and *yeeP*) [Zhaou and Rudd, 2013]. A homologous recombination between the prophage element and bacterial genomic sequences can lead to bacterial genome rearrangements [Brüssow and Hendrix, 2002] and it also leads to the gain of virulence in bacteria [Canchaya et al., 2004]. Prophages also contribute to the diversification of the bacterial genome architecture [Brüssow et al., 2004]. An integrative and comparative approach is needed to explore the evolution of pathogens and prophages to understand the gain of virulence in bacteria [Brüssow and Hendrix, 2002]. For these reasons, the evolutionary correlation of the cp4-44 prophage element with pathogenicity was studied.

The result indicates that the cp4-44 prophage element was independently gained later on the phylogeny than pathogenicity which was inherited from the common ancestor. It also indicates that pathogens are more capable of gaining the cp4-44 prophage element. The prophage elements might played a role in gaining the virulent function in bacterial pathogens [Canchaya et al., 2004], but the cp4-44 prophage element did not play a role in the raise of pathogenic characteristics in the *E. coli* + *Shigella* dataset. As prophage elements have the dynamic ability to rearrange bacterial genomic sequences [Brüssow and Hendrix, 2002], a capability of bacterial pathogens to gain prophage elements might increased the chance of virulent functions gain.

Enzyme correlated with the living environment of *Rhodobacteraceae*

(S)-2-haloacid dehalogenase (EC 3.8.1.2) is an enzyme from the hydrolase family that catalyzes the reaction "(S)-2-haloacid + H_2O = (R)-2-hydroxyacid + halide" [Schomburg et al., 2013]. Marine algae and polychaete tube worms produce diverse halogenated compounds like terpenes (pharmaceutically interesting compound) and toxic compounds (for the defense mechanism of algae). The symbiotic microorganisms live with tube worms on the surface of algae may use dehalogenase enzymes to degrade these toxic compounds. The abundant occurrence of halogenated compounds might caused marine symbiotic microbes of tube worms/marine algae to synthesize haloacid dehalogenase [Novak et al., 2013]. A haloacid dehalogenase activity tested positive for the *Rhodobacteraceae* family [Huang et al., 2011]. As this symbiotic relationship was observed in marine environmental habitats, the evolutionary correlation between (S)-2-haloacid dehalogenase activity and the living environments of *Rhodobacteraceae* spp. was hypothesized and studied.

The result indicates that the evolution of (S)-2-haloacid dehalogenase and the evolution of marine *Rhodobacteraceae spp.* were highly correlated. Both were inherited from the common ancestor of 108 genomes and lost at the same ancestral nodes. In the halogenated compound rich environment, bacteria need a defense mechanism to survive. In this manner, the (S)-2-haloacid dehalogenase is a functionally necessary enzyme for marine *Rhodobacteraceae* spp. that live along with the tube worms/algae in the marine environment [Novak et al., 2013]. Thus, this environmental relation supports the reported correlation of (S)-2-haloacid dehalogenase with the marine *Rhodobacteraceae* spp. in this study.

Pathways observed only in marine organisms and correlated with the living environment of *Rhodobacteraceae*

Ectoine **(**$C_6H_{10}N_2O_2$**)** is an osmolyte, a naturally occurring compound in halophilic microbes that increases survival chances during high osmotic stress from the environment, in particular from salt concentration and resistance towards temperature stress. Ectoine was first identified in halophilic microorganisms such as *Ectothiorhodospira halochloris, Marinococcus halophilus*, *Pseudomonas stutzeri*, and *Halomonas elongata* [Hastings et al., 2013].

The result indicates that the ectoine biosynthesis pathway was correlated in evolution with the living environment and observed only in specific marine strains of *Rhodobacteraceae* spp. Despite the fact that osmolytes in marine bacteria are the basic requirement for survival in the marine environment, ectoine biosynthesis was not inherited from the common ancestor of *Rhodobacteraceae* spp. This

suggests to study 88 marine *Rhodobacteraceae* spp. in two subsets containing 34 (presence of ectoine biosynthesis) and 54 (absence of ectoine biosynthesis) organisms respectively, so as to understand the driving mechanism for gaining ectoine as an osmoprotectant. The reason behind the ectoine presence in extant organisms that independently gained ectoine biosynthesis pathway can be explored further. The result suggests that experimental scientists to study the unknown osmoregulatory mechanism and osmoregulators in other subset containing 54 marine organisms.

Pathways observed only from non-marine organisms and correlated with the living environment of *Rhodobacteraceae*

Siderophores are low molecular weight ferric ion specific chelating compound observed in microbes. They are used by microbes to solubilize, capture and deliver ferric ion into cells. Ferric ion is present in several biological processes like oxygen transport, respiration and amino acid synthesis [Caspi et al., 2008]. Siderophores are observed more often in terrestrial microbes than marine microbes due to poor iron availability in the marine environment [Sandy and Butler, 2009]. 2,3–dihydroxy benzoate is a major iron-coordinating functional group of siderophores [Crosa and Walsch, 2002].

The result indicates that the 2,3-dihydroxy benzoate biosynthesis pathway was correlated in evolution with living environment and inherited in non-marine genomes from the common ancestral node of non-marine groups on the phylogeny. The inference supports the high dependence of the pathway on the non-marine living environment. As the marine environment is a poor source of iron, siderophores are functionally dependent on the iron-rich non-marine living environment. The study of 2,3-dihydroxy benzoate biosynthesis can be focused into the *Paracoccus* spp. subset because of their superior capability of encoding 2,3-dihydroxy benzoate biosynthesis pathways over other non-marine *Rhodobacteraceae*.

Pathways correlated with the living environment and observed in both marine and non-marine organisms in distributed manner

Ethanol degradation pathway IV is a component of the energy metabolism. Ethanol degradation pathways are classified based on the encoding enzymes or mechanism by which ethanol is oxidized to acetaldehyde. The pathways are classified from I to IV. Ethanol degradation pathway IV follows the peroxidation process in converting ethanol into acetaldehyde. The peroxidation reaction is described as “ethanol + hydrogen peroxide = aldehyde + water”. It is catalyzed by the peroxidase enzyme [Caspi et al., 2008]. Ethanol degradation pathways II and IV were observed in the *Rhodobacteraceae* dataset.

It was inferred from the result that ethanol degradation pathway IV was correlated in evolution with the living environment of *Rhodobacteraceae* spp. and several evolutionary events were observed on the phylogeny. It was distributed in 60 extant organisms. But, ethanol degradation pathway II (only difference as redox instead of peroxidation reaction) was observed in 107 out of 108 organisms as a basic ethanol metabolism. The uncorrelated evolution was observed between the ethanol degradation pathway II and living environment of *Rhodobacteraceae* spp. In comparing ethanol degradation pathway II and IV, the peroxidase enzyme (involves only in the pathway IV) was suspected as a reason for the evolutionary correlation of ethanol degradation pathway IV with the living environment. Thus, the peroxidase enzyme was cross checked for its correlation with the living environment in evolution. The study showed that the peroxidase enzyme was correlated with the living environment (p-value 0.008 and distributed in 70 extant organisms). Thus, the peroxidation process catalyzed by peroxidase enzyme could be the major driving force behind the evolutionary correlation of ethanol degradation pathway IV with the living environment of *Rhodobacteraceae*.

Pathways uncorrelated with the living environment of *Rhodobacteraceae*

Glycogen is a polysaccharide that contains only glucose. It is found in bacteria and higher organisms. It is used as an energy-storing molecule [Preiss, 1984]. Glycogen is synthesized when there is an excess of carbohydrate energy sources. Glycogen biosynthesis pathway I encodes the conversion of ADP-Glucose into glycogen [Caspi et al., 2008]. It is used during starvation time for maintaining cell growth, and during processes such as motility, pH maintenance and osmotic pressure maintenance [Preiss, 1984]. It can be inferred from its role in biological systems that it is a metabolic pathway in all glucose energy utilizing marine and non-marine organisms.

The phylogenetic tree and presence/absence of characters in a dataset were used to find correlation between characters (Pagel's method using BayesTraits software). It investigates ancestral states and identifies the probable temporal ordering of changes in two characters instead of just accounting shared inheritance. In other hand, the correlation between characters can be studied without phylogenetic tree. But, the shared phylogenetic inheritance of a pair of non-linked characters in closely related species and a pair of characters gained/lost independently in same species without evolutionary dependency could be calculated as spurious correlations [Barker and Pagel, 2005]. Only, the approach which uses the phylogenetic tree can be used to test the dependency of one character on another [Pagel, 1994] [Barker and Pagel, 2005]. The glycogen biosynthesis pathway I is an example to describe the difference

between the above approaches. Using the correlation study (chi-squared-test) without phylogenetic tree, the spurious evolutionary correlation (p-value < 0.05) was observed between the glycogen biosynthesis pathway I and the living environments of *Rhodobacteraceae* spp. But, the phylogeny (Pagel's method) based method followed in this thesis observed the uncorrelated evolution of the pathway with the living environments of *Rhodobacteraceae* spp. The character state reconstruction shows the distributed character history pattern of the pathway and the inheritance was not observed only in marine habitats.

As conclusion, a pipeline to compute evolutionary correlation between genes, enzymes, pathways and genomic features was developed and used to study evolutionary correlations specific datasets. The effect of genes lost in parallel from small genomes in the clustering process were identified and small genomes were removed from the study. Specific functional networks (e.g. motility pathways) were analysed with regard to their correlated gene pairs and verified with previous references. The correlated enzyme pairs in each pathway of the *Rhodobacteraceae* dataset were identified and applications were explained for specific pathways. The evolutionary correlation of pathways/genomic features were identified for the *Rhodobacteraceae* dataset and their applications were explained for the specific set of pathways/genomic features. The evolutionary correlation of genes with pathogenesis, correlation of enzyme, pathways with marine lifestyle were studied further by character reconstruction study and discussed. Especially, five different character history patterns were shown and discussed to understand the capability of whole genome level correlation identification approach between characters using Pagel's method for microbial genome datasets. The previously described theory/hypothesis on correlated evolution of functionally linked characters can be proved/disproved with this new approach. This study can be further extended to identify the evolutionary correlation between morphological characteristics, biochemical features, biological processes and unknown genetic elements. These strategies and implemented pipeline can be further applied for specific genomic/functional features, morphological characteristics, and unknown genetic elements. The correlated evolution between above mentioned features/characteristics can also be identified.

5. COMPUTATIONAL TOOLS AND RESOURCES

5.1. PROGRAMMING LANGUAGES

5.1.1. RUBY

Ruby is an open source, object-oriented programming language. The version Ruby 1.9.2p320 was used in this work. The Ruby is a multiple programming paradigm supported language. It supports functional, object-oriented and imperative programming. It has dynamic type system and automatic memory management. The pipeline base for the three chapters were built using the Ruby programming language. Rubymine was used as an integrated development environment for the Ruby programming language.

5.1.2. R

R is a free and open source programming language for the statistical computing and graphics. The source code of R is written in the programming languages C and FORTRAN. R is an implementation of S programming language with the semantic lexical scopes. The statistical analysis like Pearson's chi-squared test, multiple regression analysis, heat map and histograms were tested/computed using R. Rstudio was used as an integrated development environment for the R programming language.

5.2. OPERATING SYSTEM

Ubuntu is used in this work. Ubuntu is a Linux kernel based Debian distribution system. It is distributed under the free/open source software license GNU GPL. Ubuntu is built as more secure environment. It has the feature “sudo” which provides temporary rights to access all administrative tasks. The complete root access to the operating system is restricted to increase the security. Ubuntu is a supportive environment to use tools like MCL, BayesTraits software, Ruby and PostgreSQL in a single platform. Ubuntu version 10.04 and 12.04 were used in this thesis work.

5.3. CODING STANDARDS

The method of identifying the evolutionary correlation between discrete/continuous characters using BayesTraits was implemented in the Ruby programming language according to coding standards. Ruby functions with the features of “One to many characters” and “all to all characters” for correlation identification were implemented for various BayesTraits options. The coding standards were

collectively compiled by Markus Göker from the literature [Thomas et al., 2005] [Brown, 2009]. The coding standards include the Ruby programming style and project structure regulations. The Ruby programming style includes formatting, naming, documenting, testing and Ruby way of coding [Fulton, 2007]. Establishing the project structure includes the structure of files, directories and usage of external executables and libraries in the Ruby project [Thomas et al., 2005].

The formatting standards include usage of blank lines, space, blocks, operators, maximum length of a line and the parenthesis. The naming standards include English orthography, upper/lower cases used in methods/modules and Ruby functions. The script was documented in rdoc format. The DRY (Don't Repeat Yourself) [Thomas et al., 2005] principle, an extend of libraries/modules without writing new coding scripts to fulfill requirements, reduced/no use of global variables and logically split methods/modules were implemented to reduce the repetitive scripts in the Ruby code. The unit testing was implemented to ensure that old and new Ruby functions work well in the case of code expansion [Brown, 2009].

5.3. COMPUTATIONAL RESOURCES

5.3.1. DESKTOP COMPUTER

Memory – 7.7 GB

Processor – Pentium(R) Dual-Core CPU E5300 @ 2.60GHz X 2

Architecture – 64 bit

5.3.2. SERVER COMPUTER

Memory – 49.5 GB

Processor - Intel(R) Xeon(R) CPU E5420 @ 2.50GHz X 8

Architecture – 64 bit7. LIST OF TABLES AND FIGURES

6. *REFERENCES*

1. Abt, Birte, et al. "Complete genome sequence of the termite hindgut bacterium *Spirochaeta coccoides* type strain (SPN1^T), reclassification in the genus *Sphaerochaeta* as *Sphaerochaeta* coccoides comb. nov. and emendations of the family *Spirochaetaceae* and the genus *Sphaerochaeta*." *Standards in Genomic Sciences* 6.2 (2012): 194-209.

2. Abt, Birte, et al. "Genome sequence of the thermophilic fresh-water bacterium *Spirochaeta caldaria* type strain (H1^T), reclassification of *Spirochaeta caldaria* and *Spirochaeta stenostrepta* in the genus *Treponema* as *Treponema caldaria* comb. nov., *Treponema stenostrepta* comb. nov., and *Treponema zuelzerae* comb. nov., and emendation of the genus *Treponema*." *Standards in Genomic Sciences* 8.1 (2013): 88-105.

3. Anderson, Iain, et al. "Novel insights into the diversity of catabolic metabolism from ten haloarchaeal genomes." *PLoS One* 6.5 (2011): e20237.

4. Ankeny, Rachel A. "Sequencing the genome from nematode to human: changing methods, changing science." *Endeavour* 27.2 (2003): 87-92.

5. Auch, Alexander F, et al. "Genome BLAST distance phylogenies inferred from whole plastid and whole mitochondrion genome sequences."*BMC Bioinformatics* 7.1 (2006): 350.

6. Auch, Alexander F, Stefan R. Henz, and Markus Göker. "Phylogenies from whole genomes: Methodological update within a distance-based framework." *German Conference on Bioinformatics* (2008).

7. Baechle, Michael, and Paul Kirchberg. "Ruby on Rails." *IEEE Software* 24.6 (2007): 105-108.
8. Bailey, Robert E. "Global hexachlorobenzene emissions." *Chemosphere* 43.2 (2001): 167-182.
9. Baldauf, Sandra L, et al. "A kingdom-level phylogeny of eukaryotes based on combined protein data." *Science* 290.5493 (2000): 972-977.

10. Bapteste, Eric, et al. "The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba." *Proceedings of the National Academy of Sciences* 99.3 (2002): 1414-1419.

11. Barbash, Shahar, and Hermona Soreq. "Statistically invalid classification of high throughput gene expression data." *Scientific Reports* 3 (2013).

12. Barbour, Alan G, and Stanley F. Hayes. "Biology of *Borrelia* sp." *Microbiological Reviews* 50.4 (1986): 381-400.

13. Barker, Daniel, and Mark Pagel. "Predicting functional gene links from phylogenetic-statistical analyses of whole genomes." *PLoS Computational Biology* 1.1 (2005): e3.

14. Benson, Dennis A, et al. "GenBank." *Nucleic Acids Research* 28.1 (2000): 15-18.

15. Brown, Gregory T. "Ruby best practices." *O'Reilly Media, Inc.* (2009).

16. Brüssow, Harald, and Roger W. Hendrix. "Phage genomics: small is beautiful." *Cell* 108.1 (2002): 13-16.

17. Bodenreider, Olivier. "The unified medical language system (UMLS): integrating biomedical terminology." *Nucleic Acids Research* 32.suppl 1 (2004): D267-D270.

18. Brüssow, Harald, Carlos Canchaya, and Wolf-Dietrich Hardt. "Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion." *Microbiology and Molecular Biology Reviews* 68.3 (2004): 560-602.

19. Buchan, A, and José M. Gonzalez. "*Roseobacter.*" *Handbook of Hydrocarbon and Lipid Microbiology* (2010): 1335-1343.

20. Caspi, Ron, et al. "The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases." *Nucleic Acids Research* 36.suppl 1 (2008): D623-D631.

21. McRoy, Susan W, et al. "Yag: a template-based natural language generator for real-time systems." *Natural Language and Knowledge Representation Research Group, University of Wisconsin-Milwaukee* (1999).

22. Ciccarelli, Francesca D, et al. "Toward automatic reconstruction of a highly resolved tree of life." *Science* 311.5765 (2006): 1283-1287.

23. Canchaya, Carlos, Ghislain Fournous, and Harald Brüssow. "The impact of prophages on bacterial chromosomes." *Molecular Microbiology* 53.1 (2004): 9-18.

24. Codd, Edgar F. "A relational model of data for large shared data banks." *Pioneers and Their Contributions to Software Engineering* (2001): 61-98.

25. Crosa, Jorge H, and Christopher T. Walsh. "Genetics and assembly line enzymology of

siderophore biosynthesis in bacteria." *Microbiology and Molecular Biology Reviews* 66.2 (2002): 223-249.

26. Darwin, Charles. "On the origins of species by means of natural selection." *London: Murray* (1859).
27. Dayhoff, Margaret O, and Robert M. Schwartz. "Atlas of Protein Sequence and Structure." *National Biomedical Research Foundation, Washington, DC* (1978): 345-352.
28. Delsuc, Frédéric, Henner Brinkmann, and Hervé Philippe. "Phylogenomics and the reconstruction of the tree of life." *Nature Reviews Genetics* 6.5 (2005): 361-375.
29. Desper, Richard, and Olivier Gascuel. "Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle." *Journal of Computational Biology* 9.5 (2002): 687-705.
30. Dress, Andreas WM, et al. "Noisy: Identification of problematic columns in multiple sequence alignments." *Algorithms for Molecular Biology* 3.7 (2008).
31. Driskell, Amy C, et al. "Prospects for building the tree of life from large sequence databases." *Science* 306.5699 (2004): 1172-1174.
32. Droege, Stefan, et al. "*Spirochaeta coccoides* sp. nov., a Novel Coccoid Spirochete from the Hindgut of the Termite *Neotermes castaneus*." *Applied and Environmental Microbiology* 72.1 (2006): 392-397.
33. Eisen, Jonathan A. "Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis." *Genome Research* 8.3 (1998): 163-167.
34. Eisen, Jonathan A, and Claire M. Fraser. "Phylogenomics: intersection of evolution and genomics." *Science* 300.5626 (2003): 1706-1707.
35. Euzeby, J. P. "List of Prokaryotic names with Standing in Nomenclature (LPSN)." *Online at: http://www. bacterio. cict. fr/(accessed January 2010)* (2010).
36. Faraway, Julian J. "Practical regression and ANOVA using R." *Online at: http://cran.r-project.org/doc/contrib/Faraway-PRA.pdf* (2002): 125-129.
37. Felsenstein, Joseph. "Phylogenies from molecular sequences: inference and reliability." *Annual Review of Genetics* 22.1 (1988): 521-565.

38. Felsenstein, Joseph. “Inferring phylogenies. Vol. 2.” *Sunderland: Sinauer Associates* (2004).

39. Felsenstein, Joseph. “PHYLIP (Phylogeny Inference Package) version 3.6.” *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle* (2005).

40. Fielding, Roy T, and Richard N. Taylor. "Principled design of the modern Web architecture." *ACM Transactions on Internet Technology (TOIT)* 2.2 (2002): 115-150.

41. Fleischmann, Robert D, et al. "Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd." *Science* 269.5223 (1995): 496-512.

42. Fracek Jr, S. P, and J. F. Stolz. "*Spirochaeta bajacaliforniensis* sp. n. from a microbial mat community at Laguna Figueroa, Baja California Norte, Mexico." *Archives of Microbiology* 142.4 (1985): 317-325.

43. Fujibuchi, W, et al. "DBGET/LinkDB: an integrated database retrieval system." *Pac. Symp. Biocomput*. 98 (1998).

44. Fulton, Hal. “The Ruby way second edition.” *Addison-Wesley, Upper Saddle River, NJ* (2007).

45. Garrity, George M. "The State of Standards in Genomic Sciences." *Standards in Genomic Sciences* 5.3 (2011): 262-268.

46. Garrity, George M, Dawn Field, and Nikos C. Kyrpides. "Standards in Genomic Sciences." *Standards in Genomic Sciences* 1.1 (2009): 1-2.

47. Garrity, George M, et al. "Toward a standards-compliant genomic and metagenomic publication record." *OMICS A Journal of Integrative Biology* 12.2 (2008): 157-160.

48. Gascuel, Olivier. "BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data." *Molecular Biology and Evolution* 14.7 (1997): 685-695.

49. Gascuel, Olivier. "Concerning the NJ algorithm and its unweighted version, UNJ." *Mathematical Hierarchies and Biology* 37 (1997): 149-171.

50. Gerstein, Mark B, et al. "What is a gene, post-ENCODE? History and updated definition." *Genome Research* 17.6 (2007): 669-681.

51. Guindon, Stéphane, and Olivier Gascuel. "Efficient biased estimation of evolutionary distances when substitution rates vary across sites." *Molecular Biology and Evolution* 19.4 (2002): 534-

543.

52. Harris, M. A, et al. "The Gene Ontology (GO) database and informatics resource." *Nucleic Acids Research* 32. Database Issue (2004): D258-61.

53. Harvey, P. H, and Mark Pagel, “The Comparative Method in Evolutionary Biology." *Oxford University Press* (1991).

54. Hastings, Janna, et al. "The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013." *Nucleic Acids Research* 41.D1 (2013): D456-D463.

55. Hennig, Willi. "Phylogenetic systematics." *Annual Review of Entomology* 10.1 (1965): 97-116.

56. Henz, Stefan R, et al. "Whole-genome prokaryotic phylogeny." *Bioinformatics* 21.10 (2005): 2329-2335.

57. Holder, Mark, and Paul O. Lewis. "Phylogeny estimation: traditional and Bayesian approaches." *Nature Reviews Genetics* 4.4 (2003): 275-284.

58. Holland, Barbara R, et al. "δ plots: A tool for analyzing phylogenetic distance data." *Molecular Biology and Evolution* 19.12 (2002): 2051-2059.

59. Horner, David Stephen, et al. "Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing." *Briefings in Bioinformatics* 11.2 (2010): 181-197.

60. Hong, Soon Ho, Tae Yong Kim, and Sang Yup Lee. "Phylogenetic analysis based on genome-scale metabolic pathway reaction content." *Applied Microbiology and Biotechnology* 65.2 (2004): 203-210.

61. Huang, Jianyu, et al. "Phylogenetic diversity and characterization of 2-haloacid degrading bacteria from the marine sponge *Hymeniacidon perlevis*." *World Journal of Microbiology and Biotechnology* 27.8 (2011): 1787-1794.

62. Hugenholtz, Philip, Brett M. Goebel, and Norman R. Pace. "Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity." *Journal of Bacteriology* 180.18 (1998): 4765-4774.

63. Huynen, Martijn A, and Peer Bork. "Measuring genome evolution." *Proceedings of the National Academy of Sciences of the United States of America* 95.11 (1998): 5849-5856.

64. Ivanova, Natalia, et al. "Complete genome sequence of the extremely halophilic *Halanaerobium praevalens* type strain (GSL^T)." *Standards in Genomic Sciences* 4.3 (2011): 312-321.

65. Jensen, Roy A. "Orthologs and paralogs-we need to get it right." *Genome Biology* 2.8 (2001): 1002.1-1002-3.

66. Lapage, Stephen P, et al. "International code of nomenclature of bacteria: bacteriological code." *1990 revision*. *ASM Press* (1992).

67. Lester, James C, and Bruce W. Porter. "Developing and empirically evaluating robust explanation generators: The KNIGHT experiments." *Computational Linguistics* 23.1 (1997): 65-101.

68. Jill Harrison, C, and Jane A. Langdale. "A step by step guide to phylogeny reconstruction." *The Plant Journal* 45.4 (2006): 561-572.

69. Kanehisa, Minoru, and Susumu Goto. "KEGG: kyoto encyclopedia of genes and genomes." *Nucleic Acids Research* 28.1 (2000): 27-30.

70. Kanehisa, Minoru, et al. "KEGG for representation and analysis of molecular networks involving diseases and drugs." *Nucleic Acids Research* 38.suppl 1 (2010): D355-D360.

71. Kimura, Motoo. "A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences." *Journal of Molecular Evolution* 16.2 (1980): 111-120.

72. Koonin, Eugene V. "Computational genomics." *Current Biology* 11.5 (2001): R155-R158.

73. Kumar, Sudhir, and Joel Dudley. "Bioinformatics software for biologists in the genomics era." *Bioinformatics* 23.14 (2007): 1713-1717.

74. Legendre, Pierre, and Louis Legendre. "Numerical Ecology: second English edition." *Developments in Environmental Modelling* 20 (1998).

75. Lin, Jimmy, and Mark Gerstein. "Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels."*Genome Research* 10.6 (2000): 808-818.

76. Liu, Hongfang, et al. "BioThesaurus: a web-based thesaurus of protein and gene names."

Bioinformatics 22.1 (2006): 103-105.

77. Luscombe, Nicholas M, Dov Greenbaum, and Mark Gerstein. "What is bioinformatics? An introduction and overview." *Yearbook of Medical Informatics* 1 (2001): 83-99.

78. Maddison, David R. "Phylogenetic methods for inferring the evolutionary history and processes of change in discretely valued characters." *Annual Review of Entomology 39.1* (1994): 267-292.

79. Maddison, Wayne P, and David R Maddison. "Mesquite: a modular system for evolutionary analysis." (2001): Version 2.75 http://mesquiteproject.org

80. Magrane, Michele. "UniProt Knowledgebase: a hub of integrated protein data." *Database: The Journal of Biological Databases and Curation* 2011 (2011): bar009.

81. Markowitz, Victor M, et al. "The integrated microbial genomes (IMG) system." *Nucleic Acids Research* 34.suppl 1 (2006): D344-D348.

82. Medini, Duccio, et al. "The microbial pan-genome." *Current Opinion in Genetics & Development* 15.6 (2005): 589-594.

83. Meusemann, Karen, et al. "A phylogenomic approach to resolve the arthropod tree of life." *Molecular Biology and Evolution* 27.11 (2010): 2451-2464.

84. Miller, Wilmer J. “Appropriate Gene Symbols in Teaching Genetics”. *The Proceedings of the Iowa Academy of Science 92.3* (1985): 115-118.

85. Novak, Halina R, et al. "Marine Rhodobacteraceae l-haloacid dehalogenase contains a novel His/Glu dyad that could activate the catalytic water." *FEBS Journal* 280.7 (2013): 1664-1680.

86. Pagel, Mark. "Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters." *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255.1342 (1994): 37-45.

87. Pagel, Mark, Andrew Meade, and Daniel Barker. "Bayesian estimation of ancestral character states on phylogenies." *Systematic Biology* 53.5 (2004): 673-684.

88. Pagani, Ioanna, et al. "Complete genome sequence of *Desulfobulbus propionicus* type strain (1pr3^{T})." *Standards in Genomic Sciences* 4.1 (2011): 100-110.

89. Pagani, Ioanna, et al. "The Genomes OnLine Database (GOLD) v. 4: status of genomic and

metagenomic projects and their associated metadata." *Nucleic Acids Research* 40.D1 (2012): D571-D579.

90. Pati, Amrita, et al. "Complete genome sequence of *Cellulophaga lytica* type strain (LIM-21^T)." *Standards in Genomic Sciences* 4.2 (2011): 221-232.

91. Pazos, Florencio, and Alfonso Valencia. "Protein co-evolution, co-adaptation and interactions." *The EMBO journal* 27.20 (2008): 2648-2655.

92. Pearson, Karl. "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling." *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (1900): 157-175.

93. Pearson, Karl. "The problem of the random walk." *Nature* 72.1865 (1905): 294.

94. Pearson, Helen. "Genetics: what is a gene?" *Nature* 441.7092 (2006): 398-401.

95. Pellegrini, Matteo, et al. "Assigning protein functions by comparative genome analysis: protein phylogenetic profiles." *Proceedings of the National Academy of Sciences of the United States of America* 96.8 (1999): 4285-4288.

96. Philippe, Hervé, and Mathieu Blanchette. "Overview of the first phylogenomics conference." *BMC Evolutionary Biology* 7.Suppl 1 (2007): S1.

97. Pitluck, Sam, et al. "Complete genome sequence of *Thermosediminibacter oceani* type strain (JW/IW-1228P^T)." *Standards in Genomic Sciences* 3.2 (2010): 108-116.

98. Preiss, J. "Bacterial glycogen synthesis and its regulation." *Annual Reviews in Microbiology* 38.1 (1984): 419-458.

99. Quester, Susanne, and Dietmar Schomburg. "EnzymeDetector: an integrated enzyme function prediction tool and database." *BMC Bioinformatics* 12.1 (2011): 376.

100. Reenskaug, Trygve. "Models-views-controllers." *Technical Note, Xerox PARC* 32 (1979): 55.

101. Reiter, Ehud, and Robert Dale. "Building applied natural language generation systems." *Natural Language Engineering* 3.1 (1997): 57-87.

102. Ritalahti, Kirsti M, et al. "*Sphaerochaeta globosa* gen. nov., sp. nov. and *Sphaerochaeta*

pleomorpha sp. nov., free-living, spherical spirochaetes." *International Journal of Systematic and Evolutionary Microbiology* 62.1 (2012): 210-216.

103. Robinson, D. F, and Leslie R. Foulds. "Comparison of phylogenetic trees." *Mathematical Biosciences* 53.1 (1981): 131-147.

104. Rodionov, Dmitry A, Mikhail S. Gelfand, and Nicole Hugouvieux-Cotte-Pattat. "Comparative genomics of the KdgR regulon in Erwinia chrysanthemi 3937 and other gamma-proteobacteria." *Microbiology* 150.11 (2004): 3571-3590.

105. Rokas, Antonis, and Peter WH Holland. "Rare genomic changes as a tool for phylogenetics." *Trends in Ecology & Evolution* 15.11 (2000): 454-459.

106. Rubinfeld, Daniel L. "Reference guide on multiple regression." *Reference Manual on Scientific Evidence* 179 (2000).

107. Saitou, Naruya, and Masatoshi Nei. "The neighbor-joining method: a new method for reconstructing phylogenetic trees." *Molecular Biology and Evolution* 4.4 (1987): 406-425.

108. Sanderson, Michael J, et al. "Obtaining maximal concatenated phylogenetic data sets from large sequence databases." *Molecular Biology and Evolution* 20.7 (2003): 1036-1042.

109. Sandy, Moriah, and Alison Butler. "Microbial iron acquisition: marine and terrestrial siderophores." *Chemical reviews* 109.10 (2009): 4580-4595.

110. Sanger, F. "Nobel lecture: Determination of nucleotide sequences in DNA". *Nobelprize.org* (1980): Retrieved 2010-10-18.

111. Sanger, Frederick, Steven Nicklen, and Alan R. Coulson. "DNA sequencing with chain-terminating inhibitors." *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (1977): 5463-5467.

112. Schmidt, Steffen, et al. "Metabolites: a helping hand for pathway evolution?" *Trends in Biochemical Sciences* 28.6 (2003): 336-341.

113. Scholz, Matthew B., Chien-Chi Lo, and Patrick SG Chain. "Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis." *Current Opinion in Biotechnology* 23.1 (2012): 9-15.

114. Schomburg, Ida, et al. "BRENDA in 2013: integrated reactions, kinetic data, enzyme function

data, improved disease classification: new options and contents in BRENDA." *Nucleic Acids Research* 41.D1 (2013): D764-D772.

115. Shendure, Jay, and Hanlee Ji. "Next-generation DNA sequencing." *Nature Biotechnology* 26.10 (2008): 1135-1145.

116. Snel, Berend, Peer Bork, and Martijn A. Huynen. "Genome phylogeny based on gene content." *Nature Genetics* 21.1 (1999): 108-110.

117. Spring, Stefan, et al. "The Genome Sequence of *Methanohalophilus mahii* SLP^T Reveals Differences in the Energy Metabolism among Members of the *Methanosarcinaceae* Inhabiting Freshwater and Saline Environments." *Archaea* 2010 (2010): Article ID 690737, 16 pages.

118. Tatusov, Roman L, et al. "The COG database: an updated version includes eukaryotes." *BMC Bioinformatics* 4.1 (2003): 4-41.

119. Tatusov, Roman L, Eugene V. Koonin, and David J. Lipman. "A genomic perspective on protein families." *Science* 278.5338 (1997): 631-637.

120. Taylor, Chris F, et al. "Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project." *Nature Biotechnology* 26.8 (2008): 889-896.

121. Thomas, D, et al. "Programming Ruby." *The Pragmatic Bookshelf* (2005).

122. Thomas, David and David Heinemeier Hansson. "Agile web development with Rails." *The Pragmatic Bookshelf* (2006).

123. Thompson, Paul, et al. "The BioLexicon: a large-scale terminological resource for biomedical text mining." *BMC Bioinformatics* 12.1 (2011): 397.

124. Van Deemter, Kees, Emiel Krahmer, and Mariët Theune. "Real versus template-based natural language generation: A false opposition?" *Computational Linguistics* 31.1 (2005): 15-24.

125. van Dongen, Stijn Marinus. "Graph clustering by flow simulation." *PhD thesis, Universiteit Ulrecht* (2000).

126. Weisburg, William G, et al. "16S ribosomal DNA amplification for phylogenetic study." *Journal of Bacteriology* 173.2 (1991): 697-703.

127. Woese, Carl R. "Bacterial evolution." *Microbiological Reviews* 51.2 (1987): 221.

128. Wolf, Yuri I, et al. "Genome trees constructed using five different approaches suggest new major bacterial clades." *BMC Evolutionary Biology* 1.1 (2001): 8.

129. Wu, Martin, and Jonathan A. Eisen. "A simple, fast, and accurate method of phylogenomic inference." *Genome Biology* 9.10 (2008): R151.

130. Wu, Dongying, et al. "A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea." *Nature* 462.7276 (2009): 1056-1060.

131. Yasawong, Montri, et al. "Complete genome sequence of *Arcanobacterium haemolyticum* type strain (11018^{T})." *Standards in Genomic Sciences* 3.2 (2010): 126-135.

132. Zhou, Jindan, and Kenneth E. Rudd. "EcoGene 3.0." *Nucleic Acids Research 41* (2013): D613-D624.

7. ABBREVIATIONS

COG	Cluster of Orthologous Groups
GEBA	Genomic Encyclopedia of Bacteria and Archaea
GO	Gene Ontology
IMG	Integrated Microbial Genome
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Ortholog
LPSN	List of Prokaryotic Names with Standing in Nomenclature
LR	Likelihood Ratio
MCL	Markov Cluster Algorithm
MVC	Model-View-Controller
NLG	Natural Language Generation
PcoA	Principal coordinates analysis
REST	Representational State Transfer
RF	Robinson-Foulds distance
SIGS	*Standards in Genomic Sciences*
UniProtKB	UniProt Knowledge Base

8. SUPPLEMENTARY MATERIALS

Table S1 – List of organisms and corresponding dataset swith their taxa indices

Taxa index	Organism name	Dataset
t1	*Blastopirellula marina* DSM 3645	*Planctomycetes*
t2	*Planctomyces maris* DSM 8797	*Planctomycetes*
t3	*Rhodopirellula baltica* SH 1	*Planctomycetes*
t4	*Planctomyces limnophilus* DSM 3776	*Planctomycetes*
t5	*Pirellula staleyi* DSM 6068	*Planctomycetes*
t6	*Isosphaera pallida* ATCC 43644	*Planctomycetes*
t7	*Planctomyces brasiliensis* DSM 5305	*Planctomycetes*
t1	*Archaeoglobus fulgidus* DSM 4304	*Archaeoglobi* + outgroups
t2	*Methanococcus maripaludis* S2	*Archaeoglobi* + outgroups
t3	*Methanococcus maripaludis* C5	*Archaeoglobi* + outgroups
t4	*Methanococcus vannielii* SB	*Archaeoglobi* + outgroups
t5	*Methanococcus aeolicus* Nankai-3	*Archaeoglobi* + outgroups
t6	*Methanococcus maripaludis* C7	*Archaeoglobi* + outgroups
t7	*Methanococcus maripaludis* C6	*Archaeoglobi* + outgroups
t8	*Methanocaldococcus fervens* AG86	*Archaeoglobi* + outgroups
t9	*Methanocaldococcus vulcanius* M7	*Archaeoglobi* + outgroups
t10	*Archaeoglobus profundus* DSM 5631	*Archaeoglobi* + outgroups
t11	*Ferroglobus placidus* DSM 10642	*Archaeoglobi* + outgroups
t12	*Methanocaldococcus infernus* ME	*Archaeoglobi* + outgroups
t13	*Methanococcus voltae* A3	*Archaeoglobi* + outgroups
t14	*Archaeoglobus veneficus* SNP6	*Archaeoglobi* + outgroups
t15	*Methanocaldococcus jannaschii* DSM 2661	*Archaeoglobi* + outgroups
t1	*Escherichia coli* SE15	*E. coli* + *Shigella*
t2	*Escherichia coli* DH1	*E. coli* + *Shigella*
t3	*Shigella flexneri* 2002017	*E. coli* + *Shigella*
t4	*Escherichia coli* ABU 83972	*E. coli* + *Shigella*
t5	*Escherichia coli* O83:H1 str. NRG 857C	*E. coli* + *Shigella*
t6	*Escherichia coli* IHE3034	*E. coli* + *Shigella*
t7	*Escherichia coli* UM146	*E. coli* + *Shigella*
t8	*Escherichia coli* O127:H6 str. E2348/69	*E. coli* + *Shigella*
t9	*Escherichia coli* 536	*E. coli* + *Shigella*
t10	*Escherichia coli* 55989	*E. coli* + *Shigella*
t11	*Escherichia coli* APEC O1	*E. coli* + *Shigella*
t12	*Escherichia coli* BL21(DE3)	*E. coli* + *Shigella*
t13	*Escherichia coli* BW2952	*E. coli* + *Shigella*

t14	*Escherichia coli* B str. REL606	*E. coli* + *Shigella*
t15	*Escherichia coli* CFT073	*E. coli* + *Shigella*
t16	*Escherichia coli* ATCC 8739	*E. coli* + *Shigella*
t17	*Escherichia coli* E24377A	*E. coli* + *Shigella*
t18	*Escherichia coli* ED1a	*E. coli* + *Shigella*
t19	*Escherichia coli* HS	*E. coli* + *Shigella*
t20	*Escherichia coli* IAI1	*E. coli* + *Shigella*
t21	*Escherichia coli* IAI39	*E. coli* + *Shigella*
t22	*Escherichia coli* str. K-12 substr. DH10B	*E. coli* + *Shigella*
t23	*Escherichia coli* str. K-12 substr. MG1655	*E. coli* + *Shigella*
t24	*Escherichia coli* str. K-12 substr. W3110	*E. coli* + *Shigella*
t25	*Escherichia coli* O157:H7 str. Sakai	*E. coli* + *Shigella*
t26	*Escherichia coli* O157:H7 EDL933	*E. coli* + *Shigella*
t27	*Escherichia coli* O157:H7 str. EC4115	*E. coli* + *Shigella*
t28	*Escherichia coli* O157:H7 str. TW14359	*E. coli* + *Shigella*
t29	*Escherichia coli* S88	*E. coli* + *Shigella*
t30	*Escherichia coli* SE11	*E. coli* + *Shigella*
t31	*Escherichia coli* SMS-3-5	*E. coli* + *Shigella*
t32	*Escherichia coli* UMN026	*E. coli* + *Shigella*
t33	*Escherichia coli* UTI89	*E. coli* + *Shigella*
t34	*Escherichia fergusonii* ATCC 35469	*E. coli* + *Shigella*
t35	*Escherichia coli* K-12	*E. coli* + *Shigella*
t36	*Escherichia coli* ETEC H10407	*E. coli* + *Shigella*
t37	*Escherichia coli* O103:H2 str. 12009	*E. coli* + *Shigella*
t38	*Escherichia coli* O26:H11 str. 11368	*E. coli* + *Shigella*
t39	*Escherichia coli* O111:H- str. 11128	*E. coli* + *Shigella*
t40	*Escherichia coli* O55:H7 str. CB9615	*E. coli* + *Shigella*
t41	*Shigella boydii* CDC 3083-94	*E. coli* + *Shigella*
t42	*Shigella boydii* Sb227	*E. coli* + *Shigella*
t43	*Shigella dysenteriae* Sd197	*E. coli* + *Shigella*
t44	*Shigella flexneri* 2a str. 301	*E. coli* + *Shigella*
t45	*Shigella flexneri* 2a str. 2457T	*E. coli* + *Shigella*
t46	*Shigella flexneri* 5 str. 8401	*E. coli* + *Shigella*
t47	*Shigella sonnei* Ss046	*E. coli* + *Shigella*
t1	*Haloarcula marismortui* ATCC 43049	*Halobacteriales* + outgroups
t2	*Halobacterium* sp. NRC-1	*Halobacteriales* + outgroups
t3	*Haloferax volcanii* DS2	*Halobacteriales* + outgroups
t4	*Halogeometricum borinquense* PR3, DSM 11551	*Halobacteriales* + outgroups
t5	*Halomicrobium mukohataei* DSM 12286	*Halobacteriales* + outgroups
t6	*Haloquadratum walsbyi* DSM 16790	*Halobacteriales* + outgroups
t7	*Halorhabdus utahensis* DSM 12940	*Halobacteriales* + outgroups

t8	*Halorubrum lacusprofundi* ATCC 49239	*Halobacteriales* + outgroups
t9	*Haloterrigena turkmenica* VKM, DSM 5511	*Halobacteriales* + outgroups
t10	*Natronomonas pharaonis* DSM 2160	*Halobacteriales* + outgroups
t11	*Methanococcoides burtonii* DSM 6242	*Halobacteriales* + outgroups
t12	*Methanocorpusculum labreanum* Z	*Halobacteriales* + outgroups
t13	*Methanoculleus marisnigri* JR1	*Halobacteriales* + outgroups
t14	*Methanohalophilus mahii* DSM 05219	*Halobacteriales* + outgroups
t15	*Methanosaeta thermophila* PT	*Halobacteriales* + outgroups
t16	*Methanosarcina acetivorans* C2A	*Halobacteriales* + outgroups
t17	*Methanosphaerula palustris* E1-9c	*Halobacteriales* + outgroups
t18	*Methanospirillum hungatei* JF-1	*Halobacteriales* + outgroups
t19	*Methanocella paludicola* SANAE	*Halobacteriales* + outgroups
t1	*Bacteroides thetaiotaomicron* VPI-5482	*Bacteroidales*
t2	*Porphyromonas gingivalis* ATCC 33277	*Bacteroidales*
t3	*Bacteroides vulgatus* ATCC 8482	*Bacteroidales*
t4	*Parabacteroides distasonis* ATCC 8503	*Bacteroidales*
t5	*Prevotella melaninogenica* ATCC 25845	*Bacteroidales*
t6	*Paludibacter propionicigenes* WB4	*Bacteroidales*
t7	*Bacteroides helcogenes* P 36-108	*Bacteroidales*
t8	*Bacteroides salanitronis* DSM 18170	*Bacteroidales*
t9	*Odoribacter splanchnicus* DSM 20712	*Bacteroidales*
t10	*Porphyromonas asaccharolytica* DSM 20707	*Bacteroidales*
t11	*Bacteroides fragilis* NCTC 9343	*Bacteroidales*
t12	*Alistipes shahii* WAL 8301	*Bacteroidales*
t13	*Bacteroides xylanisolvens* XB1A	*Bacteroidales*
t1	*Oceanicola batsensis* HTCC2597	Roseobacter clade
t2	*Oceanicola granulosus* HTCC2516	Roseobacter clade
t3	*Pelagibaca bermudensis* HTCC2601	Roseobacter clade
t4	*Oceanibulbus indolifex* HEL-45	Roseobacter clade
t5	*Phaeobacter gallaeciensis* BS107	Roseobacter clade
t6	*Roseobacter denitrificans* OCh 114	Roseobacter clade
t7	*Ruegeria pomeroyi* DSS-3	Roseobacter clade
t8	*Dinoroseobacter shibae* DFL 12	Roseobacter clade
t9	*Roseobacter litoralis* Och 149	Roseobacter clade
t1	*Borrelia valaisiana* VS116	*Spirochaetae*
t2	*Borrelia burgdorferi* B31	*Spirochaetae*
t3	*Treponema denticola* ATCC 35405	*Spirochaetae*
t4	*Spirochaeta thermophila* DSM 6578	*Spirochaetae*
t5	*Treponema azotonutricium* ZAS-9	*Spirochaetae*

t6	*Treponema primitia* ZAS-2	*Spirochaetae*
t7	*Spirochaeta smaragdinae* DSM 11293	*Spirochaetae*
t8	*Treponema succinifaciens* DSM 2489	*Spirochaetae*
t9	*Sphaerochaeta coccoides* DSM 17374	*Spirochaetae*
t10	*Treponema brennaborense* DSM 12168	*Spirochaetae*
t1	*Acidothermus cellulolyticus* 11B	*Actinomycetales*
t2	*Actinosynnema mirum DSM* 43827	*Actinomycetales*
t3	*Beutenbergia cavernae* HKI 0122, DSM 12333	*Actinomycetales*
t4	*Clavibacter michiganensis sepedonicus* ATCC 33113	*Actinomycetales*
t5	*Corynebacterium efficiens* YS-314	*Actinomycetales*
t6	*Corynebacterium glutamicum* ATCC 13032 (Bielefeld)	*Actinomycetales*
t7	*Corynebacterium glutamicum* ATCC 13032 (Kitasato)	*Actinomycetales*
t8	*Jonesia denitrificans* DSM 20603	*Actinomycetales*
t9	*Kineococcus radiotolerans* SRS30216	*Actinomycetales*
t10	*Kribbella flavida* DSM 17836	*Actinomycetales*
t11	*Kytococcus sedentarius* DSM 20547	*Actinomycetales*
t12	*Mycobacterium tuberculosis* H37Rv	*Actinomycetales*
t13	*Mycobacterium vanbaalenii* PYR-1	*Actinomycetales*
t14	*Renibacterium salmoninarum* ATCC 33209	*Actinomycetales*
t15	*Saccharomonospora viridis* DSM 43017	*Actinomycetales*
t16	*Salinispora tropica* CNB-440	*Actinomycetales*
t17	*Sanguibacter keddieii* ST-74, DSM 10542	*Actinomycetales*
t18	*Streptomyces avermitilis* MA-4680	*Actinomycetales*
t19	*Thermobispora bispora* R51, DSM 43833	*Actinomycetales*
t20	*Thermomonospora curvata* DSM 43183	*Actinomycetales*
t21	*Tropheryma whipplei Twist*	*Actinomycetales*
t22	*Xylanimonas cellulosilytica* XIL07, DSM 15894	*Actinomycetales*
t23	*Acidimicrobium ferrooxidans* DSM 10331	*Actinomycetales*
t24	*Arcanobacterium haemolyticum* CCM, DSM 20595	*Actinomycetales*
t25	*Atopobium parvulum* DSM 20469	*Actinomycetales*
t26	*Bifidobacterium adolescentis* ATCC 15703	*Actinomycetales*
t27	*Bifidobacterium animalis subsp. lactis* DSM 10140	*Actinomycetales*
t28	*Bifidobacterium dentium* Bd1	*Actinomycetales*
t29	*Bifidobacterium longum infantis* ATCC 15697	*Actinomycetales*
t30	*Catenulispora acidiphila* DSM 44928	*Actinomycetales*
t31	*Conexibacter woesei* DSM 14684	*Actinomycetales*
t32	*Corynebacterium aurimucosum* ATCC 700975	*Actinomycetales*
t33	*Corynebacterium kroppenstedtii* DSM 44385	*Actinomycetales*
t34	*Cryptobacterium curtum* DSM 15641	*Actinomycetales*
t35	*Eggerthella lenta* DSM 2243	*Actinomycetales*
t36	*Geodermatophilus obscurus* DSM 43160	*Actinomycetales*

t37	*Gordonia bronchialis* DSM 43247	*Actinomycetales*
t38	*Kocuria rhizophila* DC2201	*Actinomycetales*
t39	*Micrococcus luteus* NCTC 2665	*Actinomycetales*
t40	*Mycobacterium abscessus*	*Actinomycetales*
t41	*Nocardiopsis dassonvillei dassonvillei* DSM 43111	*Actinomycetales*
t42	*Olsenella uli* VPI, DSM 7084	*Actinomycetales*
t43	*Segniliparus rotundus* DSM 44985	*Actinomycetales*
t44	*Slackia heliotrinireducens* DSM 20476	*Actinomycetales*
t45	*Stackebrandtia nassauensis* LLR-40K-21, DSM 44728	*Actinomycetales*
t46	*Tsukamurella paurometabola* DSM 20162	*Actinomycetales*
t47	*Arthrobacter chlorophenolicus* A6	*Actinomycetales*
t48	*Brachybacterium faecium* DSM 4810	*Actinomycetales*
t49	*Corynebacterium urealyticum* DSM 7109	*Actinomycetales*
t50	*Nakamurella multipartita* DSM 44233	*Actinomycetales*
t51	*Streptosporangium roseum* NI 9100, DSM 43021	*Actinomycetales*
t52	*Rubrobacter xylanophilus* DSM 9941	*Actinomycetales*
t53	*Saccharopolyspora erythraea* NRRL 2338	*Actinomycetales*
t54	*Cellulomonas flavigena* 134, DSM 20109	*Actinomycetales*

Table S2 – List of organisms from Spirochaetae 29 dataset with taxa indices

Taxa index	**Organism name**
t3	*Treponema denticola* ATCC 35405
t5	*Treponema azotonutricium* ZAS 9
t6	*Treponema primitia* ZAS 2
t7	*Spirochaeta smaragdinae* DSM 11293
t9	*Sphaerochaeta coccoides* DSM 17374
t10	*Treponema brennaborense* DSM 12168
t11	*Sphaerochaeta globosa*
t12	*Sphaerochaeta pleomorpha* str. Grapes
t13	*Leptospira biflexa* serovar Patoc strain Patoc 1 Paris
t14	*Brachyspira murdochii* DSM 12563
t15	*Spirochaeta caldaria* DSM 7334
t16	*Brachyspira intermedia* PWS A
t17	*Spirochaeta thermophila* DSM 6578
t20	*Treponema succinifaciens* DSM 2489
t21	*Treponema saccharophilum* DSM 2985
t23	*Spirochaeta africana* DSM 8902
t24	*Leptonema illini* DSM 21528
t25	*Brachyspira hyodysenteriae* ATCC 27164
t26	*Brachyspira innocens* ATCC 29796
t27	*Brachyspira pilosicoli* P43 6 78
t28	*Leptospira alexanderi* serovar Manhao 3 str L 60
t29	*Leptospira broomii* serovar Hurstbridge str 5399
t30	*Leptospira inadai* serovar Lyme str 10
t31	*Leptospira kmetyi* serovar Malaysia str Bejo Iso9
t32	*Leptospira licerasiae* serovar Varillal str VAR 010
t33	*Leptospira santarosai* serovar Shermani str LT 821
t34	*Spirochaeta alkalica* DSM 8900
t35	*Spirochaeta bajacaliforniensis* DSM 16054
t36	*Turneriella parva* DSM 21527

Table S3 – List of organisms from Rhodobacteraceae 108 dataset with taxa indices

Taxa index	**Organism name**
t1	*Oceanicola batsensis* HTCC2597
t2	*Oceanicola granulosus* HTCC2516
t3	*Pelagibaca bermudensis* HTCC2601
t4	*Oceanibulbus indolifex* HEL-45
t7	*Ruegeria pomeroyi* DSS-3
t9	*Roseobacter litoralis* OCh 149
t10	*Labrenzia aggregata* IAM
t12	*Leisingera aquimarina* DSM 24565
t14	*Leisingera nanhaiensis* NH52F
t15	*Marinovum algicola* DSM 10251
t17	*Octadecabacter antarcticus* 307
t18	*Octadecabacter arcticus* 238
t19	*Phaeobacter arcticus* DSM 23566
t20	*Phaeobacter caeruleus* 13
t22	*Phaeobacter gallaeciensis* CIP 105210
t23	*Phaeobacter inhibens* T5
t24	*Rhodobacter sphaeroides* 2.4.1
t28	*Ruegeria lacuscaerulensis* ITI-1157
t29	*Sagittula stellata* E-37
t31	*Jannaschia* sp. CCS1
t32	*Ruegeria* sp. TM1040
t33	*Paracoccus denitrificans* PD1222
t34	*Rhodobacter sphaeroides* ATCC 17029
t35	*Rhodobacter sphaeroides* ATCC 17025
t36	*Rhodobacter sphaeroides* KD131
t37	*Rhodobacter capsulatus* SB 1003
t38	*Ketogulonicigenium vulgare* WSH-001
t39	*Ketogulonicigenium vulgare* Y25
t40	*Polymorphum gilvum* SL003B-26A1
t41	*Phaeobacter gallaeciensis* 210
t42	*Phaeobacter gallaeciensis* DSM 17395

t43	*Pseudovibrio* sp. FO-BEG1
t46	*Loktanella vestfoldensis* SKA53
t48	*Roseobacter* sp. MED193
t50	*Rhodobacteraceae bacterium* HTCC2150
t51	*Roseobacter* sp. CCS2
t52	*Roseobacter* sp. SK209-2-6
t53	*Roseovarius* sp. TM1035
t54	*Roseobacter* sp. AzwK-3b
t55	*Rhodobacter* sp. SW2
t56	*Ahrensia* sp. R2A130
t58	*Citreicella* sp. 357
t59	*Rhodovulum* sp. PH10
t62	*Citreicella* sp. SE45
t63	*Oceanicola* sp. S124
t64	*Paracoccus* sp. N5
t65	*Paracoccus* sp. TRP
t66	*Pseudovibrio* sp. JE062
t67	*Rhodobacteraceae bacterium* HTCC2083
t68	*Rhodobacteraceae bacterium* KLH11
t69	*Rhodobacterales bacterium* Y4I
t70	*Roseibium* sp. TrichSKD4
t71	*Roseobacter* sp. GAI101
t72	*Ruegeria* sp. R11
t73	*Ruegeria* sp. TW15
t74	*Silicibacter* sp. TrichCH4B
t75	*Thalassiobium* sp. R2A62
t76	*Ahrensia kielensis* DSM 5890
t77	*Roseobacter denitrificans* OCh 114
t78	*Dinoroseobacter shibae* DFL 12
t79	*Labrenzia alexandrii* DFL-11
t80	*Leisingera methylohalidivorans* MB2
t81	*Loktanella hongkongensis* DSM 17492
t82	*Marinovum algicola* DG 898

t83	*Maritimibacter alkaliphilus* HTCC2654
t84	*Paracoccus denitrificans* SD1
t85	*Phaeobacter daeponensis* TF-218
t86	*Rhodobacter sphaeroides* WS8N
t87	*Rhodobacterales bacterium* HTCC2255
t89	*Roseovarius nubinhibens* ISM
t90	*Roseovarius* sp. 217
t91	*Roseovarius mucosus* DSM 17069
t92	*Rubellimicrobium mesophilum* DSM 19309
t93	*Rubellimicrobium thermophilum* DSM 16684
t94	*Salipiger mucosus* DSM 16094
t95	*Sulfitobacter* sp. EE-36
t96	*Sulfitobacter* sp. NAS-14.1
t97	*Thalassobacter arenae* DSM 19593
t99	*Paracoccus* sp. J55
t100	*Rhodobacter* sp. AKP1
t101	*Celeribacter baekdonensis* B30
t102	*Loktanella* sp. SE62
t103	*Oceaniovalibus guishaninsula* JLT2003
t104	*Phaeobacter gallaeciensis* ANG1
t105	*Roseibacterium elongatum* OCh 323
t106	*Roseobacter* sp. R2A57
t107	*Wenxinia marina* DSM 24838

Table S4 – List of genes co-involve in motility pathways of Spirochaeta dataset with 29 genomes

Gene/Character index	Gene/character names
c408	30s ribosomal protein s2
c22	Response regulator receiver modulated CheB methylesterase
c3666	Nucleoside-diphosphate-sugar epimerase
c190	NAD(P)-dependent iron-only hydrogenase diaphorase component flavoprotein
c5757	Septum formation initiator
c5565	Hypothetical protein
c5769	Conserved hypothetical protein
c5730	HipA n-terminal domain protein
c4373	Hypothetical protein
c4890	Glycosyl transferase family 2
c3036	Arsr family transcriptional regulator
c2924	Outer membrane protein
c1536	Flagellar export protein fliJ
c5299	Hypothetical protein
c424	Fatty acid hydroxylase
c477	30s ribosomal protein s14
c495	1-deoxy-d-xylulose-5-phosphate synthase
c526	PhoH family protein
c527	Aldose 1-epimerase
c105	Aldo/keto reductase
c479	Electron transport complex, RnfABCDGE type, C subunit
c129	Glutamate synthase (NADPH), homotetrameric
c493	Chromosome segregation protein SMC
c461	Extracellular solute-binding protein family 1
c507	1-acyl-sn-glycerol-3-phosphate acetyltransferase
c463	GntR domain protein
c5614	Hypothetical protein
c5663	Hypothetical protein

c2975	Lipoprotein
c3446	LicD family protein
c3945	Magnesium chelatase, subunit D/I family
c4110	Transcriptional regulator, AsnC family
c4056	DNA adenine methylase

Table S5 – List of genomic features and their abbreviations used. The proportion of those genes in the genome are included in the second column.

Abbreviations	**Genomic feature**
J	Translation proteins
A	RNA processing and modification proteins
K	Transcription proteins
L	Replication and repair proteins
B	Proteins involve in chromatin structure and dynamics
D	Proteins involve in cell cycle control and mitosis
Y	Proteins involve in nuclear structural arrangements
V	Proteins involve in defense mechanism
T	Proteins involve in signal transduction
M	Proteins involve in cell wall membrane biogenesis
N	Proteins involve in cell motility
Z	Proteins in cytoskeleton
W	Proteins in extracellular structures
U	Proteins involve in intracellular trafficing and secretion
O	Proteins involve in post translational modifications
C	Proteins involve in energy production and conversion
G	Proteins involve in carbohydrate metabolism and transport
E	Proteins involve in amino acids metabolism and transport
F	Proteins involve in nucleotide metabolism and

	transport
H	Proteins involve in co-enzyme metabolism
I	Proteins involve in lipid metabolism
P	Proteins involve in inorganic ion transport and metabolism
Q	Proteins involve in secondary metabolite biosynthesis, transport and catabolism
R	Proteins of general functional predictions
S	Proteins with functions unknown
SSU	Number of small subunit rRNA in a genome
LSU	Number of large subunit rRNA in a genome
CRISPR	Number of Clustered Regularly Inter spaced Short Palindromic Repeats in a genome
tRNA	Number of tRNA in a genome
ribo.prot	Number of ribosomal proteins in a proteome
transport	Number of transporter proteins in a proteome
ABC	Number of ATP Bining Cassettes in a genome
phage	Number of phage like protein elements found in a genome

Table S6 – List of pathways correlated in evolution with lifestyle of Rhodobacteraceae dataset with 108 genomes

Pathway ID	**Pathway name**	**P-value**	**Number of marine organisms**	**Number of non-marine organisms**
143	trans, trans-farnesyl diphosphate biosynthesis [3]	5.3209×10^{-6}	62	4
287	all-trans-farnesol biosynthesis [4]	5.3209×10^{-6}	62	4
280	glycogen biosynthesis II (from UDP-D-Glucose) [4]	5.0685×10^{-6}	10	10
183	starch biosynthesis [6]	4.2152×10^{-6}	28	13
206	mixed acid fermentation [13]	3.0281×10^{-6}	13	10
289	isoleucine biosynthesis II [6]	7.5454×10^{-5}	81	12
155	trehalose biosynthesis V [3]	6.9905×10^{-5}	10	3
198	guanosine nucleotides degradation I [4]	1.5689×10^{-5}	75	17
20	mycolate biosynthesis [6]	0.0001	9	1
291	2'-deoxy-alpha-D-ribose 1-phosphate degradation [3]	0.0002	33	11

115	respiration (anaerobic) [9]	0.0002	19	8
81	chlorophyllide a biosynthesis III [7]	0.0002	2	3
223	chlorophyllide a biosynthesis II [7]	0.0002	2	3
126	nitrate reduction II (assimilatory) [3]	0.0002	35	8
170	tyrosine degradation I [5]	0.0003	45	7
16	ornithine biosynthesis [5]	0.0004	86	15
301	arginine biosynthesis III [9]	0.0004	86	15
131	formaldehyde assimilation II (RuMP Cycle) [8]	0.0004	14	12
90	vitamin B6 salvage (plants) [4]	0.0004	10	2
279	UDP-N-acetylmuramoyl-pentapeptide biosynthesis II (lysine-containing) [8]	0.0006	83	15
197	glycogen degradation II [4]	0.0008	20	9
32	D-galactonate degradation [3]	0.0008	15	0
33	glycine betaine degradation [5]	0.0008	83	15
316	guanosine nucleotides degradation III [5]	0.0009	80	17
18	arginine biosynthesis IV [5]	0.0009	76	9
177	6-hydroxymethyl-dihydropterin diphosphate biosynthesis I [3]	0.0009	54	4
169	methylerythritol phosphate pathway [8]	0.0010	87	18
25	D-galacturonate degradation I [4]	0.0011	26	3
106	valine degradation I [7]	0.0011	6	8
187	protocatechuate degradation II (ortho-cleavage pathway) [4]	0.0014	57	10
185	lysine biosynthesis II [8]	0.0016	48	7
56	creatinine degradation II [5]	0.0017	5	6
113	glycerol degradation to butanol [10]	0.0017	8	4
160	inosine-5'-phosphate biosynthesis II [5]	0.0019	85	14
196	unsaturated, even numbered fatty acid beta-oxidation [5]	0.0020	23	0
228	flavin biosynthesis III (fungi) [7]	0.0021	42	2
148	formaldehyde oxidation II (glutathione-dependent) [3]	0.0023	7	12
283	pyruvate fermentation to acetate IV [3]	0.0023	4	2
243	urea cycle [5]	0.0025	75	16
258	ethanol degradation IV [3]	0.0026	43	17
285	isoleucine biosynthesis IV [6]	0.0027	78	12
248	NAD/NADH phosphorylation and dephosphorylation [4]	0.0028	67	15
239	formaldehyde assimilation I (serine pathway) [10]	0.0031	15	6
219	glycine cleavage [3]	0.0032	82	19

27	thiamin salvage II [3]	0.0035	11	2
2	C4 photosynthetic carbon assimilation cycle, NAD-ME type [6]	0.0036	14	6
39	putrescine biosynthesis IV [4]	0.0038	66	14
312	TCA cycle IV (2-oxoglutarate decarboxylase) [9]	0.0039	72	11
207	ribitol degradation [3]	0.0044	12	10
128	docosahexanoate biosynthesis I [4]	0.0044	12	5
21	gluconeogenesis II (Methanobacterium thermoautotrophicum) [12]	0.0045	17	8
260	superoxide radicals degradation [3]	0.0045	36	14
221	beta-D-glucuronide and D-glucuronate degradation [5]	0.0045	26	5
314	glycolysis IV (plant cytosol) [10]	0.0047	24	12
71	ketolysis [3]	0.0049	85	17
151	S-adenosyl-L-methionine cycle II [3]	0.0049	40	5
246	D-galactarate degradation II [3]	0.0050	8	2
129	ectoine biosynthesis [5]	0.0051	34	0
124	coenzyme A biosynthesis [4]	0.0051	88	17
300	5-aminoimidazole ribonucleotide biosynthesis I [5]	0.0051	88	17
9	ppGpp biosynthesis [4]	0.0053	88	17
165	UTP and CTP dephosphorylation I [4]	0.0066	29	2
179	methylmalonyl pathway [3]	0.0067	87	16
83	glycolysis II (from fructose-6P) [10]	0.0069	63	18
200	CDP-diacylglycerol biosynthesis I [4]	0.0069	69	13
286	CDP-diacylglycerol biosynthesis II [4]	0.0069	69	13
168	tetrahydrofolate biosynthesis [3]	0.0070	39	3
158	myo-inositol degradation [4]	0.0070	46	10
74	nitrate reduction VII (denitrification) [4]	0.0071	28	13
256	pyridine nucleotide cycling (plants) [8]	0.0071	41	3
211	2-oxopentenoate degradation [3]	0.0076	0	1
218	biphenyl degradation [4]	0.0076	0	1
295	sulfate reduction II (assimilatory) [3]	0.0077	44	12
204	acetylene degradation [5]	0.0078	21	9
8	palmitate biosynthesis II (bacteria and plants) [8]	0.0081	1	0
53	palmitate biosynthesis I (animals and fungi) [8]	0.0081	1	0
322	ethylene biosynthesis V [6]	0.0083	16	6
132	pyruvate fermentation to butanoate [6]	0.0089	3	1
152	Entner-Doudoroff pathway II (non-phosphorylative) [8]	0.0089	3	1

87	pyruvate fermentation to butanol I [5]	0.0090	3	1
307	glutamate degradation VII (to butanoate) [3]	0.0090	3	1
203	chlorosalicylate degradation [3]	0.0090	3	2
161	serine biosynthesis [3]	0.0090	87	17
141	isoleucine degradation I [5]	0.0091	36	9
6	pyruvate fermentation to ethanol I [3]	0.0091	3	1
249	sulfate reduction III (assimilatory) [3]	0.0109	16	0
167	fatty acid beta-oxidation I [6]	0.0110	87	16
60	tRNA processing [5]	0.0114	86	16
102	4-hydroxybenzoate biosynthesis V [5]	0.0114	24	2
118	tetrapyrrole biosynthesis I (from glutamate) [6]	0.0118	19	3
44	4-hydroxyproline degradation II [3]	0.0122	23	0
116	uracil degradation I (reductive) [3]	0.0129	71	9
241	thymine degradation [3]	0.0129	71	9
99	methylaspartate cycle [13]	0.0130	78	13
125	pyrimidine deoxyribonucleotides de novo biosynthesis I [9]	0.0132	14	0
234	Entner-Doudoroff pathway III (semi-phosphorylative) [10]	0.0134	18	3
304	superpathway of fermentation (Chlamydomonas reinhardtii) [7]	0.0134	3	3
7	ketogenesis [5]	0.0135	10	2
107	salicortin biosynthesis [5]	0.0136	1	1
119	NAD biosynthesis I (from aspartate) [6]	0.0150	49	11
253	C4 photosynthetic carbon assimilation cycle, NADP-ME type [4]	0.0152	25	8
202	pyruvate fermentation to ethanol III [3]	0.0160	0	1
57	chlorophyllide a biosynthesis I [7]	0.0162	7	1
261	adenosylcobalamin biosynthesis II (late cobalt incorporation) [9]	0.0165	53	12
37	adenosylcobalamin salvage from cobinamide II [7]	0.0182	0	2
137	seleno-amino acid biosynthesis [4]	0.0188	88	18
130	photorespiration [6]	0.0192	4	3
292	5-dehydro-4-deoxy-D-glucuronate degradation [4]	0.0194	18	5
122	lysine biosynthesis III [6]	0.0196	2	2
12	tryptophan degradation X (mammalian, via tryptamine) [4]	0.0202	7	2
34	methionine biosynthesis I [5]	0.0202	80	16
172	formaldehyde assimilation III (dihydroxyacetone cycle) [11]	0.0207	65	19

35	2,3-dihydroxybenzoate biosynthesis [3]	0.0222	0	8
147	pyridoxal 5'-phosphate biosynthesis I [6]	0.0229	1	3
233	starch degradation V [5]	0.0247	16	9
303	Calvin-Benson-Bassham cycle [11]	0.0269	14	11
61	glycolysis VI (mammalian) [10]	0.0269	88	18
72	acetyl-CoA fermentation to butyrate II [5]	0.0281	77	14
306	glutamine biosynthesis III [7]	0.0281	25	6
180	2-methylcitrate cycle I [5]	0.0293	5	3
98	glyoxylate cycle [4]	0.0298	84	18
30	ubiquinol-9 biosynthesis (prokaryotic) [4]	0.0301	72	12
209	ubiquinol-8 biosynthesis (prokaryotic) [4]	0.0301	72	12
236	ubiquinol-7 biosynthesis (prokaryotic) [4]	0.0301	72	12
274	ubiquinol-10 biosynthesis (prokaryotic) [4]	0.0301	72	12
91	nitrifier denitrification [4]	0.0309	30	12
278	sulfolactate degradation II [3]	0.0320	32	0
10	incomplete reductive TCA cycle [7]	0.0323	9	3
290	cis-dodecenoyl biosynthesis [6]	0.0326	2	0
262	catechol degradation to beta-ketoadipate [4]	0.0338	19	6
78	fatty acid beta-oxidation II (peroxisome) [5]	0.0340	87	17
294	alanine biosynthesis I [3]	0.0348	25	1
217	leucine degradation I [5]	0.0361	79	15
66	C4 photosynthetic carbon assimilation cycle, PEPCK type [7]	0.0372	1	2
144	nicotine biosynthesis [3]	0.0372	45	10
117	proline biosynthesis I [3]	0.0381	85	18
42	ubiquinol-9 biosynthesis (eukaryotic) [4]	0.0381	71	12
43	ubiquinol-7 biosynthesis (eukaryotic) [4]	0.0381	71	12
138	ubiquinol-6 biosynthesis (eukaryotic) [4]	0.0381	71	12
215	ubiquinol-10 biosynthesis (eukaryotic) [4]	0.0381	71	12
194	tryptophan degradation to 2-amino-3-carboxymuconate semialdehyde [5]	0.0405	11	3
54	atrazine degradation I (aerobic) [3]	0.0419	18	2
15	gallate degradation I [4]	0.0427	1	0
80	gallate degradation II [4]	0.0427	1	0
272	methylgallate degradation [5]	0.0427	1	0
305	gentisate degradation [3]	0.0427	1	0
230	fatty acid biosynthesis initiation I [5]	0.0437	79	16
184	purine nucleobases degradation II (anaerobic) [14]	0.0465	2	1
38	homoserine biosynthesis [3]	0.0466	87	18

265	leucine degradation III [3]	0.0467	0	1
268	3-methylbutanol biosynthesis [3]	0.0467	0	1
157	TCA cycle VII (mammalian) [7]	0.0468	16	5
134	heme biosynthesis from uroporphyrinogen-III I [4]	0.0478	86	17
231	superpathway of heme biosynthesis from uroporphyrinogen-III [4]	0.0478	86	17
28	(R)-cysteate degradation [3]	0.0496	0	2
153	tetrapyrrole biosynthesis II (from glycine) [4]	0.0499	88	19
245	pyruvate decarboxylation to acetyl CoA [3]	0.0499	88	19
271	glycolysis III (glucokinase) [10]	0.0499	88	19

9. CURRICULUM VITAE

PERSONAL DETAILS

Name: Palani Kannan Kandavel

Born on: 29.05.1985

Born in: Virudhunagar, India

Nationality: Indian

EDUCATION AND BACKGROUND

1990 – 2002 – Kshatriya Vidyasala Higher secondary school, India – Secondary and Higher secondary

2002 – 2006 – Kamaraj College of Engineering and Technology, India – Bachelor in Technology

2007 – 2008 – Anna University, India – Masters in Science (By Research)

2009 – 2010 – National Resource Centre for Free/Open Source Softwares, India – Project Engineer I

2010 – 2013 – Leibniz Institut DSMZ, Germany – Wissenschaftliche Hilfskraft

Since 2014 – Max Planck-Genome-centre Cologne, Germany – Wissenschaftlicher Mitarbeiter